! TECHNICAL GUIDE #10 !

DESIGNING METHODS FOR THE MONITORING AND EVALUATION OF HOUSEHOLD FOOD SECURITY RURAL DEVELOPMENT PROJECTS

Calogero Carletto
Saul S. Morris

International Food Policy Research Institute
2033 K Street, N.W.
Washington, D.C. 20006 U.S.A.

March, 1999

CONTENTS

1. INTRODUCTION[1]

In recent years, many development agencies have made intensive efforts to improve their efficiency and increase their impact on rural poverty. At the heart of this new strategic management process is the measurement of performance. With Household Food Security (HFS) and nutritional security now clearly identified as desired outcomes of many development projects, there is a need to assess the performance of investment projects in terms of their impact on the HFS and nutrition status of their targets groups.

When the target populations of development agencies are highly risk-prone, they require rigorous formulation and monitoring. Poorly thought-out evaluations may inadvertently act as an incentive to target better-off elements in the projects' zones of influence, who offer higher returns and promise faster disbursement of project resources. In addition, there is a clear danger of prioritizing more easily measurable outcomes or indicators, which fail to provide the information necessary to address broader objectives or to enhance the effectiveness of rural development projects for "the poorest of the poor." In addition, proper evaluations call for an increased awareness that less tangible objectives—such as the formation of social capital, for example—may pursue. Less tangible and broader development objectives do not, however, justify less rigorous evaluation methods. On the contrary, they call for subtler and more sensitive methodologies and indicators.

This guide emphasizes the design of quantitative impact evaluation exercises for HFS and nutrition, and provides development practitioners with the basic principles on why, when and how to choose and implement a particular evaluation system. We argue that two of the key features of a good impact evaluation study are the availability of accurate baseline information and a properly thought-out control group, respectively allowing for *before-after* and *with-without* comparisons. The importance of a joint temporal and cross-sectional comparison of the beneficiary group against a counterfactual is crucial to simultaneously control for the effects of all sorts of external factors likely to contaminate the impact evaluation results. We also argue that the involvement of the evaluation team in the earliest stages of project design stage is the

[1] Funding for data collection and analysis of these data has been supported by the International Fund for Agricultural Development (TA Grant No. 301-IFPRI). We gratefully acknowledge this funding, but stress that ideas and opinions presented here are our responsibility and should, in no way, be attributed to IFAD.

most suitable way to ensure a proper and accurate evaluation without having to rely on more complicated statistical techniques, as well as to permit a sound learning process to ensue from the evaluation exercise. However, if the conditions dictate, statistical techniques can still provide the evaluation team with effective tools for a well-founded impact evaluation.

In the following sections we draw on seminal work recently completed by the UNICEF Evaluation Office, in an attempt to provide the reader with the conceptual underpinnings for the choice of a particular design suited to the type and the level of accuracy of the information required by the different intended end-users. In the second part of the document, we report on two of the evaluation methodologies used in the field in the course of projects focused on strengthening HFS and nutritional aspects of poverty alleviation projects.

What kinds of information should be sought?

A comprehensive evaluation exercise can be conceived of as closely following the chronological and logical progression of a project cycle, and comprises four sequential steps: the assessment of first the *provision*, then the *utilization*, *coverage* and *impact*[2] of new services (Habicht, Victora & Vaughan, 1997). The provision of a service by a project, if extended to and properly utilized by a sufficiently large number of beneficiaries, is expected to have an impact on certain variables of interest among the beneficiary population. A number of relations and assumptions link the provision of the service to its impact. A thorough understanding of the existence and strength of these linkages will have a major effect on the types of instrumentalities proposed by the project and, ultimately, on the design of the evaluation system.

Following this rationale, the first objective of an evaluation exercise is usually to assess service provision. Once the provision of the service has been ascertained, it may be important to evaluate the level of utilization of such services by the intended beneficiaries, and their coverage (take-up) by the project's target groups. It is only when the correct service is provided in a timely

[2] By *provision* is meant availability of new services, such as credit lines with commercial banks. *Utilization* implies the measurement of the rate of use of these services, such as disbursement of loans to smallholder farmers. The issue of *coverage* asks whether the target population is being reached—for instance, what proportion of poor smallholder farmers has been able to take out a new loan?

manner and properly utilized by a sufficiently large number of beneficiaries, that one can plausibly expect an impact on the indicator of interest. Only in these cases is an *impact* evaluation required or justified.

Project evaluation can thus be seen as a gradual process leading to impact evaluation when the information is required and the conditions call for it. For example, if—based on a preliminary evaluation of the project implementation—we have been able to assert with a certain degree of confidence that the provision of the services provided by the project was largely inadequate, or that the level of utilization of the service by the targeted beneficiaries was extremely low, then the situation may not justify pursuing the evaluation further to measure impact. Even in situations in which we have been able to conclude that the project has reached a large group of beneficiaries, and the service has been widely utilized, an impact evaluation exercise may be fruitless if the project has been short-lived or the nature of the intervention is such as to make it unreasonable to expect results within the time period elapsed.

On the other hand, limiting an evaluation to an assessment of service provision, utilization or take-up, based on shaky assumptions about the relationships between project interventions and end-results can be equally improper and misleading. For instance, stopping short of measuring the impact that a small-scale irrigation project has had on the food security of the households adopting the technology, based on the simplistic assumption that improved irrigation must have had an effect on household agricultural output and access to food, is likely to be inappropriate.

While there undoubtedly are cases in which it is possible to assume that the next link is automatic (i.e., if there is provision of services, there will be utilization; if there is utilization, there will be coverage; if there is coverage, there will be impact), this "blind faith" should not be allowed to become routine and, especially for the purpose of impact evaluation, it should be exercised if and only if the nature of the association between the process and impact indicator is well proven. Most importantly, when a weak point is discovered in the chain, e.g., provision did not result into utilization, or coverage did not lead to impact, then the evaluation should include a review of the institutional design of the intervention (see Technical Guide #11) and the underlying model of the relationship between the intervention and expected impacts (for an introduction, see Technical Guide #1).

How accurate should the evaluation be?

In addition to deciding on what to measure, another issue in the design of the evaluation system is to determine the level of specificity of the information sought, as required by the different stakeholders. Again borrowing terminology from UNICEF (Habicht, Victora, and Vaughan 1997), we can identify three different types of statements reflecting different degrees of confidence end-users may require from the evaluation results: adequacy, plausibility, and probability.

An adequacy assessment simply determines whether some outcome actually occurred as expected, e.g., did food security/nutritional status improve? This type of assessment may be particularly relevant when evaluating process indicators such as the provision, utilization or coverage of a particular project activity, e.g., the distribution of improved seed varieties. It tends, however, to be of little use for impact evaluation since it is unable to isolate the effects of the project from those of other concurrent processes, e.g., whether an observed improvement in yields was due to the provision of improved seed varieties by the project, or instead could be partly or completely attributed to, say, unusually good weather in the area of project intervention. Adequacy assessments are attractive since they do not require working with a control group—for example, farmers who did not receive the improved seed. This greatly reduces the complexity of the data collection activities, but the limitations are obvious.

In contrast to adequacy assessments, plausibility assessments permit determination of whether a given change can actually be attributed to the project, by isolating its effect from all other confounding factors. In the above example, one might ask whether the improved seed program had an impact on household incomes? By disentangling the project effects from other confounding factors one can state that the project appears to have had an effect *above and beyond the impact of nonproject influences*.

The need to control for this confounding arises from the fact that over the project life cycle, it is likely that external factors may contribute, positively or negatively, to changes in outcomes measured among project participants. For example, an observed improvement in child nutritional status over the course of the project could be partly due to an inflow of humanitarian food aid increasing food availability in the area. Similarly, in the context of generalized deterioration, any measurement of project impact would tend to underestimate the true effects,

since the project activities may have served as a safety net against concurrent adversity, such as a drought or a drop in food aid. For this reason, the mere comparison of indicators before and after the project is very likely to result in misleading results, since it is based on the faulty assumption that the two time periods exhibited similar circumstances except for the presence of the project.

An indispensable feature of plausibility assessments is the use of a control group. Ideally, the control group will exhibit identical characteristics (on aggregate) to the beneficiary group, except for project participation. In reality, this is often not the case, since project participants are rarely chosen at random. There is, therefore, ample potential for project beneficiaries to exhibit characteristics that are systematically different to the control group. It follows that in addition to controlling for the external confounding factors potentially affecting the observed trends, it is imperative for a proper impact assessment to also control for this between-group heterogeneity. The potential bias likely to arise from a nonrandom participation scheme is generally referred to as *selection bias* (or, when project participation is a choice variable, *self-selection bias*).

Based on the objectives of the evaluation exercise and the constraints dictated by the specific conditions, one can select an internal or an external control group. An internal group is formed by elements in the same area of project intervention who could have joined the program but elected, or were constrained, not to do so. Alternatively, an external control group generally includes those units located in an area not served by the project for whom the option to join was never available. The control for location-specific confounding will be required in this latter case unless one can make a strong case for the assumption of homogeneity.

Finally, probability evaluations can ensure that there is a small, known probability that the differences between project and control areas were due to confounding, to a systematic bias, or to chance. The basis for such a type of statement is random allocation to project intervention or control status, allowing one to determine with a given probability that the average features of the intervention and control groups are identical. The principle of randomization may appear daunting, but in most circumstances, it is relatively straightforward. In addition to randomization, a probability statement will require the adequate statistical power; without this, a probability statement becomes simply a plausibility statement.

A rigorous plausibility or probability evaluation will usually be based on a longitudinal-control design, allowing for both before-after and with-without comparisons. The basis for a

longitudinal-control study is the access to baseline information compatible with the objective of the evaluation and the availability of a properly selected control group. Both basic requirements for either a probability or a plausibility inference, i.e., baseline information and a control group, call for the early involvement of the evaluation team, since the onset of the project activities, to influence the project development process and ensure an adequate earmarking of project resources to the evaluation activities.

In conclusion, the appropriateness of adequacy, plausibility or probability evaluation depend on a number of factors such as (1) the objectives of the project, (2) the technical skills of its implementors, (3) the identity and technical sophistication of the end-users of the evaluation results, and (4) resource and time availability. Generally, it may be desirable that new project approaches be rigorously evaluated in at least one location using probability evaluations. In more routine situations, the use of plausibility evaluations may be more cost-effective and sufficient to provide decision-makers with broad policy options based on a wide spectrum of experiences. Adequacy assessments should be avoided.

What indicators should be used in evaluation?

As has been repeatedly observed, the choice of indicators to be used in assessing project impact will depend on the stated objectives of the project and on the use to which the evaluation is to be put. This in turn will require the identification of the end-users of the information to be generated. Also, simplicity and replicability of the indicator may be important. Whenever feasible, the inclusion of both process and impact indicators in the evaluation system will allow different stakeholders to assess a project-induced change in its many dimensions. The question of selection of outcome indicators for nutrition and HFS is extensively discussed in the relevant guides (#5 and #7, respectively). Any indicator that is chosen should reflect the true and broad objectives of the project being evaluated (see Technical Guide #11 for a discussion).

Avoiding Perverse Incentives in Evaluation of Development Projects

A poorly thought-out monitoring and/or evaluation scheme is likely to create perverse incentives for project managers and implementors alike. The choice of a particular design or set of indicators will almost certainly affect the way that project activities are selected and

implemented. A well thought-out monitoring and evaluation scheme is intended to feed into the different stages of project development and contribute to the correct identification of instruments and methods. For instance, it is clear that putting an emphasis on targeting the poorest may affect traditional performance indicators, and as such the results of the project should not be belittled. For example, assessing the performance of a credit system solely on types of financial indicators such as disbursement or repayment rates is likely to create an incentive for the project management to service primarily better-off beneficiaries, who are the ones most likely to repay. The assessment of the credit program based on the disaggregation of loan disbursement rates by socioeconomic group of recipients could be a way to partly circumvent this problem. Alternatively, different weights could be assigned to the different groups, creating a bias in favor of more marginal elements. Keeping the evaluation simple by providing more accessible and timely information will allow better monitoring of project activities by project stakeholders. The call for simplicity and rigor of the evaluation system are not conflicting concepts and should be pursued in parallel. In summary, more creative approaches are needed for an improved assessment of project success/failure in light of the Fund's broader and unique mandate, and the requirements for timely and accurate information to feed into policy-makers' decision-making processes.

2. CASE STUDIES

In this section we report on the impact evaluation methodologies used in two of the case studies conducted as part of IFPRI projects in Honduras and Malawi. In Honduras, the low coverage of the program, the existence of a highly comparable, accessible yet geographically isolated control group, and the phased nature of the intervention facilitated the design of a robust "quasi-experimental" evaluation (see Valadez and Bamberger 1994). In Malawi, the lack of baseline information—combined with the nonrandom nature of project intervention—dictated a different approach. Statistical manipulation and data collection techniques based on recall methods were used to control for the potential biases and other confounding. Both evaluation methods relied on a control group and were an attempt to make strong plausibility statements about the projects' impacts on HFS and nutrition. In view of the difficulties involved and the

expertise required, the statistical approach used in the Malawi case study must generally be seen as a reserve option for development projects. Early involvement of the evaluation team and the commitment to provide solid technical and financial backing to the evaluation system can help overcome the implementation constraints of the statistical approach.

Honduras

This section describes the evaluation of the interim impact on Household Food Security and nutrition of the Rural Development Program of the Western Region (PLANDERO) in Honduras, close to the borders with Guatemala and El Salvador. The evaluation aimed to provide a strong plausibility statement of the degree to which observed changes in both "service" coverage and final outcome indicators could be attributed to project activities. This approach was chosen once it became apparent that random allocation of project interventions between beneficiary and control communities would not be politically feasible in this setting.

The evaluation design took advantage of two fortuitous features of the PLANDERO project: first, the incorporation of beneficiaries was to be phased over several years; second, the overall coverage of the project did not (and was never intended to) exceed 8 percent (5,000/60,000) of the target group of poor rural households in the area.

The phased incorporation of project beneficiaries meant that even though the evaluation started one year into the execution of the project, it was possible to identify a sample of households (clustered in producers' associations) who were just about to receive credit and technical support for the first time. This permitted the collection of "baseline" information, against which the impact of subsequent PLANDERO-related activities could be assessed. Prospective monitoring of this group alone would have been sufficient to identify changes of the type:

$$\Delta_I = I_1 - I_0,$$

where I_0 is the average HFS or nutrition status of beneficiary households just prior to receiving services or benefits for the first time (referred to as *time_0*), I_1 is the status of beneficiary households after the introduction of the intervention (*time_1*), and Δ_I is the observed change, the *difference* in average HFS/nutrition status between time_0 and time_1. For example, I_0 might be

the average dietary energy intake of beneficiary households on incorporation into the project, I_1 might be their average energy intake a year later, and Δ_I would be the difference between the two, a positive quantity where the situation improved, negative where the situation deteriorated, and equal to zero where the situation remained stable. It is, of course, understood that some of the changes identified in the beneficiary group will have been the direct result of project activities, while others are partially due to project activities and partially the result of changes in external factors, and others still are entirely due to changes in the external environment.

Another advantage of the phased incorporation of project beneficiaries in Western Honduras was the existence of a pool of communities that had already been earmarked for inclusion in PLANDERO at a later date. These communities generally had functioning farmers' associations, and many of them were already well-known to the technical assistance companies that delivered services under contract from PLANDERO. From this pool of "reserve" communities, it was possible to identify a second sample of households (also clustered in producers' associations, for the most part) that were similar to the intervention group *I*, could also be observed from time_0 to time_1, but would not over this period benefit from the technical assistance or credit available through PLANDERO. This group of communities are referred to as the *control communities*, *C*. Monitoring of this group permitted estimation of the parameters

$$\Delta_C = C_1 - C_0,$$

where C_0 is the average HFS or nutrition status of control community households at time_0, C_1 is the status of control community households at time_1, and Δ_C, the difference between these two quantities, may be interpreted as a measure of the changes due to factors *external to PLANDERO* affecting outcomes observed from time_0 to time_1.

Valadez and Bamberger (1994, 235-237) have shown that if it can additionally be assumed that (a) the control and intervention communities were similar at time_0, (b) the external factors affecting the control communities and the intervention communities are the same, and (c) the effects of the program are strictly limited to the intervention communities, *I*, then the impact of the project can be estimated as

$$\Delta_I - \Delta_C = I_1 - I_0 - C_1 - C_0.$$

In the Honduras case study, great care was taken at the design stage to ensure that the control communities were similar to the intervention communities at time_0: in fact, they were matched one-to-one on the basis of geographical area, altitude, and production system. The intervention and control communities were similar , though not identical, at time_0, as shown below.

	Intervention communities n=193	Control communities n=189
Household size (mean/s.d.)	6.6 (2.7)	6.7 (2.7)
Asset score (mean/s.d.)	2.0 (0.7)	2.0 (0.7)
Land ownership, hectares (mean/s.d.)	3.0 (4.5)	2.8 (3.5)
Total cultivated area, hectares (mean/s.d.)	2.1 (2.1)	2.0 (1.9)

Because of the geographical proximity of the control and intervention communities, most of the external factors affecting their food security status would have been uniform. For example, the unusual weather patterns (attributable to the El Niño phenomenon) observed between time_0 and time_1 were common to the entire study area. Similarly, the government's decision to import large quantities of maize in mid-1997 resulted in large drops in the maize price between the 1996-97 harvest and the subsequent one in 1997-98 (median maize prices fell in the intervention communities from L.150/quintal to L.115/quintal, and in the control communities from L.150 to L.110). This design is, however, vulnerable to idiosyncratic changes affecting single communities. It can only be hoped that by including a large number of different communities in the control sample, the net effect of the sum of all such idiosyncracies will be zero.

The fact that controls were selected from different communities from the intervention sites had the advantage that there was little contamination of the control communities by project activities (at least during the one-year period observed), which would have the undesirable effect of diluting the apparent project impact. Use of radio programs meant that a certain amount of contamination did occur: for example, even in the control communities, 6 percent of respondents heard about Integrated Pest Control through PLANDERO, and 13 percent heard about Rural

Savings Associations from this source. Nonetheless, the ability of farmers in the control communities to transform this knowledge into practice without the support of the project remained very limited.

The importance of including a control group can be seen by considering the case of dietary energy intake. From time_0 to time_1, energy intake in the intervention group increased by just 1.6 percent, a negligible change. However, over the same period, energy intake in the control communities fell by 6.0 percent, suggesting that members of the PLANDERO group may in fact have been protected from a small deterioration in energy intake affecting other households in the region. Although the net effect of the project on energy intake from time_0 to time_1 (+7.7 percent) did not attain statistical significance, it should certainly be borne in mind that were this trend to continue for the remainder of the project life cycle, the end result could be very significant, both statistically and substantively.

In order to try and compensate for the short period of observation between time_0 and time_1, the Honduras case study supplemented the basic longitudinal-control design with an additional group of households (denoted I') belonging to producers' associations that had already been receiving technical assistance and credit from PLANDERO for a whole year before the beginning of the evaluation period. Once again, these communities were individually matched to control communities and new-intervention communities on the basis of geographic location, altitude, and production system. The main purpose of including this group of households in the evaluation was to determine whether some of the differences identified between the new-intervention group and the control group could be expected to be maintained or even increased over time. For example, it was found that adoption of the agronomic practices promoted by PLANDERO increased with duration in the program (all contrasts statistically significant at the $P<0.05$ level) as shown below.

	Percentage of households reporting adoption of recommended practice		
	Control communities	1 year with PLANDERO	2 years with PLANDERO
Rural Savings Associations	39%	69%	76%
Vaccination of fowl and swine	23%	30%	45%
Correct density of seeds	78%	87%	89%
Hedges	57%	71%	75%
Integrated Pest Control	35%	44%	49%
Organic fertilizers	39%	51%	53%

It was also assumed that some of the changes resulting from project activities would take time to manifest themselves, perhaps only becoming evident after a "latent" period of intensive training and opinion forming. Thus, it was noted that between time_0 and time_1, households in the control communities increased their total dietary diversity by approximately one food item, while those in the intervention communities increased their diversity scores by 2.5 items on average, and those households in their second year with PLANDERO increased their scores by fully five items (trend statistically significant at the $P<0.05$ level).

The Honduras case study was fortunate that the matching procedure employed appeared to result in broadly comparable intervention and control groups at time_0. However, there is always a concern that the intervention and control groups may, in fact, have been on different trajectories prior to time_0, intersecting only temporarily at this time; thus, what appeared to be a project impact at time_1 could, in fact, have been nothing more than the inevitable crossing of preexisting trajectories in the control and intervention communities. This topic is dealt with extensively in Valadez and Bamberger (1994, 245-246), and requires monitoring over a longer period of time than was possible in Honduras to convincingly purge the results of the "burden of history." Worse, still, the two groups could be similar on the *n* variables observed at time_0, but quite different on scores of others that were not or could not be observed. As mentioned previously, one way of avoiding this problem is to assign communities to intervention and control groups at random, a process that involves little extra effort when—as in the Honduras

case—there is a large pool of "reserves" that cannot, in any case, all be served in the first phase of project implementation. There are, however, formidable political constraints to random allocation of communities, which may, as was to be the case in Honduras, prove to be insuperable.

Malawi

The Malawi case study presented a less-than-optimal scenario from the evaluator's perspective. Among the reasons for this were that

- there was no adequate baseline study of the project area,
- project beneficiaries could not be considered a random selection of all households in the area,
- significant changes in the economic environment occurred following project inception,
- there were no up-to-date data at the village or section level to allow the identification of comparable external control areas,
- there was no up-to-date sampling frame available for nonbeneficiary households,
- time and resource constraints made it unfeasible to construct a comprehensive sampling frame.

A number of techniques were used to ensure a reliable evaluation study in the face of these numerous limitations, including

- use of recall methods to reconstruct the situation of both beneficiaries and control in the pre-project period, thus permitting "before-after" comparisons,
- the application of so-called "two-stage estimation procedures" to control for differences between beneficiaries and nonbeneficiaries, arising from the nonrandom selection of beneficiaries from the population (selection bias),
- choice of an internal control group, thus eliminating the need to obtain information on nonbeneficiary communities,

- use of EPI cluster-sampling methods to identify a representative control group in the absence of a comprehensive sample frame (see Guide #8).

One of the objectives of the Malawi case study was to assess the impact of project participation between time_0 (time of project onset) and time_1 (time of the evaluation) on a set of chosen indicators. Random allocation of the project intervention would have supported the hypothesis of homogeneity between participants and nonparticipants at time_0 by ruling out the possibility that a biased (self-) selection process was at play. However, as a result of the nonrandom allocation of project resources, project beneficiaries were likely to have exhibited characteristics at time_0 that were systematically different from the rest of the population. Thus, a straight comparison of beneficiary and control groups at time_1 would almost certainly have been biased, and would have led to misleading estimates of project impact. It was therefore imperative to control for the potential selectivity bias in the analysis.

Availability of baseline information describing the two groups in the pre-project period would have made the estimation procedure more straightforward and accurate, had the proper information been collected at time_0. In the Malawi case study, the lack of baseline information made it necessary to use recall methods at the data collection stage for both the beneficiary and control groups, to permit the before-after comparison. The possibility of constructing a "longitudinal" data set from a cross-section at time_1 depends on the length of the recall period (i.e., the time elapsed between time_0 and time_1), as well as the particular data collection techniques used, the nature of the variables of interest, and the availability of technical expertise and trained personnel in the field to elicit this type of historical information. Fortunately, the relatively recent implementation of the SFSP project much reduced the difficulties of using recall methods. The magnitude of the problem would have been much larger if the evaluation team had needed to reconstruct historical data going back several years.[3] The prior identification of household- and location-specific temporal benchmarks facilitated the work of the enumerators in

[3] Even in these cases, however, going back a few additional years in the recall of major events is recommended since they may still have an influence on the households decision at time_0. In the Malawi case study, although the project has been operational for only two years, the recall period went back up to 7 years for some variables.

assisting the respondent in the recall process. Even so, the dangers and difficulties of reconstructing historical data should not be underestimated.

In this study, the use of an external control group was not deemed appropriate. This was because of the lack of information at a spatially disaggregated level that would have allowed the identification of a control group outside of the project area with characteristics similar to the beneficiary group, eliminating the need to control in the analysis for location-specific differences between beneficiaries and controls. In addition, because of the low project coverage within the area of intervention, the main disadvantage of using an internal control group (potential project spill-over effects to nonparticipants) was considered negligible. Because of the lack of sampling frame and the need to reduce the likelihood of a biased selection of the control group, a variant of the EPI cluster-sample design was used to select the control group (for more details on the approach, please refer to sampling, Guide #8).

To illustrate the approach used in the Malawi case study and highlight the consequences of ignoring selection bias, the following paragraphs guide the reader through a fully worked out example of the estimation of the project impact on a selected indicator of nutritional status, height-for-age Z-score of children 6 - 60 months of age (for more details on this and other similar indicators, see Guide #5). For the theoretical and more technical presentation of the selection model, we refer the reader to Appendix 1.

As shown below (Table 1), a straight comparison of the prevalence of stunting between children of participant households and those from the control group suggests that there was essentially no difference in the prevalence of stunting between project children and control group children.

Table 1 Percentage of stunting (height-for-age Z-score < -2)

	Project	Control	Both
Number of observations	111	153	264
Percent stunting (HAZ<-2)	52	56	54

This initial result should in no way be interpreted as evidence of the lack of project impact. Multivariate analysis (statistical modeling) would provide more appropriate evidence by accounting in a simultaneous fashion for more than one determinant of nutritional status.

One such method often (but improperly) used is to estimate a multiple linear regression model using Ordinary Least Square (OLS) estimation methods where we include the height-for-age Z-scores as the response variable and a binary (yes/no) variable reflecting participation in the program as one of the explanatory variables, together with all other variables believed to determine child nutritional status. This type of model can be run very easily using virtually any statistical package or spreadsheet application. In Table 2, we report the estimated coefficient and the standard error of the participation variable for the Malawi case study.

Table 2 Impact of project participation (from OLS model)

	Coefficient	Standard Error
Project participation	-.022	.18

Note: Estimated coefficient after controlling for several child, mother and household characteristics (the full model estimation is reported in Table 4 in Appendix 2).

Just as in the straight comparison, the result suggests (erroneously) that project participation has no effect on child nutritional performance. Because participation in the project is not random, however, the estimate of this coefficient is inaccurate. For a household to join the project, it must satisfy a set of restrictive eligibility criteria. Yet not all households meeting these criteria would decide to join the program. This selection rule indicates that, there are both selection (by the project, based on eligibility criteria) and self-selection (by beneficiaries, who elect to enter the program based on some idiosyncratic selection rule, e.g., expected returns, or lack of alternative credit sources) processes at play that invalidate the OLS results. To account for the nonrandomness of this selection rule, we first estimate the probability for a household to join the program by regressing the participation variable on a number of regressors believed to have affected the selection rule (results of the estimation are in Appendix 2). The Probit results yield an estimated variable, the Inverse Mills Ratio (IMR), which, in broad terms, can be interpreted as a variable capturing all those unobservable characteristics potentially having an

effect on the final outcome (nutritional status), and which differentiate the two groups beyond the project effect.

In order to run the Probit model, we choose a number of variables likely to be associated to the decision to join the program, but that are uncorrelated to child nutritional status. As previously noted, being determinants of the participation decision, the variables included in this first-stage equation should reflect household circumstances at time_0, i.e., before the onset of the project activities. As applied in the example, options for the evaluation team include: (1) using post-project variables that are unlikely to have changed during the course of the project life cycle, e.g., educational level of the household head; (2) using recall methods to collect information on pre-project status that are relatively easy for the respondent to remember, e.g., household composition, sale/purchase of major assets, cropping patterns of major crops; (3) a combination of the two. Caution should be used when selecting variables observed at time_1, since there is always a risk that these may have been affected by the program. For example, while landownership at time_0 may be an appropriate choice of variable to explain program participation, the same variable at time_1 may give rise to biased estimates if participation in the program has affected household land accumulation patterns between time_0 and time_1. Examples of a good variable to use in this first stage would be whether the household head previously knew the extension workers in charge of promoting project membership, or whether any relative or friend had already joined the project. Conversely, examples of less appropriate variables would be women's education or a wealth proxy such as availability of a latrine, since although likely to reflect the household human and capital wealth (possible determinants of participation), they are both also likely to be related to children's nutritional performance. It should be noted that with these methods, failure to identify variables that correctly predict project participation will prevent estimation of project impact. Therefore choosing these variables requires carefully planning before the beginning of data collection activities, and demands familiarity both with the method and with local conditions.

The estimated IMR is then included in a second-stage equation that looks exactly like the first OLS equation, except for the added selectivity variable. This second equation can safely be estimated by OLS. The estimated coefficients of the participation and selectivity (IMR) variables are reported below in Table 3 (the full estimation results appear in Appendix 2).

Table 3 Impact of project participation (from SELECTION model)

	Coefficient	Standard Error
Project participation	1.07	.50
Selectivity bias (IMR)	-0.75	.33

Note: Estimated coefficients after controlling for several child, mother and household characteristics (the full model estimation is reported in Table 4 in Appendix 2).

A test for presence of bias in the program selection process is a test on the coefficient of the IMR. The negative significant value of the coefficient (-.75) reflects the existence of a negative selectivity bias against participants and a positive selectivity bias in favor of nonparticipants, indicating that project participants exhibit unobservable characteristics not inducive of good nutrition, and that, perhaps because of that, were purposively selected into the program.

Under the assumption that the selection model (the first stage Probit) is correctly specified, the coefficient on the participation variable (1.08) now reflects the true impact of project participation on nutritional performance. The result is quite striking: the estimated coefficient is quite large in magnitude, and strongly statistically significant. Participation in the project appears to be associated with an improvement in the height-for-age Z-score of preschoolers by one whole point (one standard deviation). The interpretation of the results is that the project was successful in targeting the worse-off households (as reflected in the negative coefficient on the selectivity variable) and, within the elapsed project cycle, in raising the nutritional performance of their preschoolers to the level comparable to the one exhibited by the control children.

In summary, we have shown a simple example in which correcting for selectivity bias has important consequences for the results. It is also clear that the methods required for this correction demand considerable technical expertise, and are unlikely to be well-suited for routine use in country.

APPENDIX 1
THE SELECTION MODEL

In this appendix, we illustrate the theory underlying the selection model and explain the rationale for its use in the context of project evaluation. Specifically, the statistical model used in the Malawi case study is generally known in economic literature as a *treatment model*. The treatment model can be seen as a variant of a general selection model in which differences in unobservable characteristics driving the selection process may account for part or all of the project impact estimated using standard (OLS) methods. In the presence of selectivity bias caused by the nonrandomness of the selection rule, the OLS coefficient for project participation is inconsistent, resulting in over/underestimation of the true contribution of the project to the outcome of interest.

Let us assume we wish to estimate the project effect for a given performance indicator Y (for more details on the choice of food and nutrition security indicators, see Guides #5 and #7) e.g., the height-for-age Z-scores of preschoolers. The flaw in the straight comparison of mean scores is that it implicitly assumes that the benefits from joining the program for participants would be the same as the benefits for nonparticipants had they chosen to join the program. Because participation is the outcome of a nonrandom process, this is unlikely to be the case; rather, it will reflect idiosyncratic household characteristics, both observable and unobservable.

One method often (but improperly) used is to estimate a multiple linear regression model using the outcome indicator (Y) as dependent variable and a binary (yes/no) variable reflecting participation in the program (C) as one of the explanatory variables, together with all other variables believed to determine child nutritional status (X):

$$Y=\beta X+\delta C+\epsilon\ . \qquad \textbf{(1)}$$

The problem with this approach is that, due to the nonrandom nature of beneficiary selection into the program, the resulting regression coefficient on the participation variable C (δ) is likely to also capture the contribution to Y of other factors, not included in the Xs, which are specific to the group of beneficiaries, yet are external to the project. For example, if the project

only attracted better-off farmers (or more highly skilled farmers, or farmers distinguished by any other set of unmeasured or unmeasurable characteristics), it is reasonable to assume that their children's nutritional status would have been better, regardless of project intervention when compared to children of nonparticipants. Conversely, if the program explicitly targets the most destitute households, its impact could be to provide a safety net preventing those households falling any further behind the nonparticipants in terms of the indicator of interest. In both cases, estimation by a standard linear regression model of the coefficient δ will over/underestimate the true project effect.

Under the assumption that these unobservable characteristics do not change over the course of the project life cycle, the availability of baseline information would permit partial control of this unobservable heterogeneity by estimating the impact over first differences. However, once again, this would only estimate the impact for the households who had joined the program, and no inference could be made regarding the control group.

An appealing and relatively straightforward alternative involves estimating the participation effect in two stages (Heckman 1979). In the first stage, we model the household decision to join the program as

$$\Gamma = \gamma W + \upsilon \,, \qquad \textbf{(2)}$$

in which an individual household will decide to join the program if and only if $\Gamma > 0$. For example, excluding the possibility of restrictive eligibility criteria, each individual household will decide to join if the expected returns from joining the project are expected to outweigh the benefits from not joining, i.e., the difference in returns is greater than zero. But Γ is not observable, so we define an observable binary variable C that will take the value of 1 if the household decides to adopt, and $C = 0$ otherwise. We then estimate the following equation using Probit (or Logit) methods:

$$\begin{array}{lll} C = \gamma' W + \upsilon & \textit{where} & \\ C = 1 & \textit{iff } \Gamma > 0 & . \\ C = 0 & \textit{otherwise} & \end{array} \qquad \textbf{(3)}$$

Assuming that C is a realization of a binomial stochastic process, we define the probability of a household joining the program as

$$Prob(C=1)=Prob(\Gamma>0)=Prob(\upsilon>-\gamma W)=1-Prob(\upsilon\leq-\gamma W)=Prob(\gamma W)=\Phi(\gamma W)\,, \qquad \textbf{(4)}$$

where Φ is the cumulative distribution of υ, assumed to be distributed normally. As we can see from expression (4), the probability of a household joining the project (i.e., of C taking the value of one) depends on the household-specific variables W and *is not random*, since households exhibiting characteristics predisposing towards better outcomes also have a higher probability of joining the program (under random selection, each household has an equal probability of selection). Consequently, the mean returns of participants will equal

$$E(Y|C=1)=E(Y|\Gamma>0)=E(Y|\upsilon>-\gamma W)=\beta'X+\delta+E(\epsilon|\upsilon>-\gamma W)=\beta'X+\delta+S=\beta'X+\Delta\,. \qquad \textbf{(5)}$$

Regressing Y on X and C would only yield

$$E(Y) \;=\; \beta'X+\delta \qquad \textbf{(6)}$$

and therefore correctly estimates the impact of the program C if and only if $\delta = \Delta$, which is when $S = 0$. But for S to equal zero, the two error terms ϵ and υ must be independent, which only hold if the selection rule is random and not a function of W.

Without going into the proof (we refer the interested reader to Greene 1993), and assuming that the error terms are distributed according to a bivariate Normal distribution with mean zero, and correlation ρ, then

$$\begin{aligned} E(\epsilon|\upsilon>-\gamma W)&=\rho\sigma[\phi(\gamma W)/\Phi(\gamma W)]=\rho\sigma\lambda_1 \quad \text{for participants, and} \\ E(\epsilon|\upsilon\leq-\gamma W)&=\rho\sigma[-\phi(\gamma W)/1-\Phi(\gamma W)]=\rho\sigma\lambda_0 \quad \text{for nonparticipants}\,, \end{aligned} \qquad \textbf{(7)}$$

where λ_i are the "derived inverse Mills ratio (IMR)," ρ is the correlation coefficient between the two error terms, and $\rho\sigma$ equals the regression coefficient on the IMR, β_λ. As shown above, the

IMR is the ratio of the value of the density function of a standard normal distribution calculated at γW and the probability of being in the sample, which equals the value of the cumulative distribution at γW for participants and its complement to 1 for nonparticipants.

Once the IMR is included in the second-stage equation, the coefficient on the participation variable δ can be interpreted with more confidence as capturing the effect of project participation on the nutritional performance of the child. The estimated coefficient on the IMR β_λ would give the magnitude and direction of the selection bias. A positive coefficient is interpreted as a selection bias in favor of participants and against nonparticipants, i.e., project beneficiaries had a higher probability of being selected into the program, or, in other words, beneficiaries exhibit unobserved (or unobservable) characteristics predisposing them to higher returns.

APPENDIX 2

In this appendix, we report the full estimation of the Malawi case study example presented in the main text. The following model is an oversimplification of the true structural model explaining nutritional status and should be taken as no more than an illustration of the selectivity model.

Table 4 Ordinary Least Squares (OLS) model without selectivity correction
Dependent variable: Height-for-age Z-score of preschoolers age 6-60 months
Number of observations: 264

	Coefficient	t	Significance
Child characteristics			
Age	-.01	-2.2	**
Sex	.05	0.32	
Mother characteristics			
Age	.03	2.80	***
Height	.03	2.13	**
Number of years of education	-.17	-2.10	**
Number of years of education squared	.02	2.46	**
Household characteristics			
Source of drinking water [(a)]	-.01	-.05	
Distance of water source [(b)]	-.29	-1.49	
Access to latrine/toilet	.33	1.50	
Participation in program	-.02	-.13	

Notes: ***(**){*} significant at 1 percent (5 percent){10 percent}.
[a] = 0 if water source is river/stream, = 1 otherwise (borehole, public or private tap).
[b] = 0 if drinking water source within 100 meters, = 1 otherwise.

Table 5 PROBIT of project participation (1st stage SELECTION model)
Dependent variable: Participation in program
Number of observations: 264

	Coefficient	Z	Significance
Household head characteristics			
Age	.004	.48	
Sex	-.24	1.22	
Education	-.05	-1.74	*
Household characteristics			
Owned land in 1995 in acres	0.10	2.39	**
Number of pre-adoption shocks	-.16	-1.81	*
Know extension worker?	1.64	2.84	***
Number of year known extension worker	-.16	-1.53	
Number of years squared	.013	2.00	**

Notes: ***(**){*} significant at 1 percent (5 percent){10 percent}.

Table 6 SELECTION model (2nd stage OLS with selectivity correction)
Dependent variable: Height-for-age Z-score
Number of observations: 264

	Coefficient	t	Significance
Child characteristics			
Age	-.01	-2.22	**
Sex	.04	.28	
Mother characteristics			
Age	.02	2.59	***
Height	.03	2.11	**
Number of years of education	-.16	-2.06	**
Number of years of education squared	.02	2.53	**
Household characteristics			
Source of drinking water [(a)]	.02	.11	
Distance of water source [(b)]	-.30	-1.55	
Access to latrine/toilet	.24	1.07	
Participation in program	1.08	2.18	**
Selectivity bias (IMR)	-.75	-2.30	**

Notes: ***(**){*} significant at 1 percent (5 percent){10 percent}.
[a] = 0 if water source is river/stream, =1 otherwise (borehole, public or private tap).
[b] = 0 if drinking water source within 100 meters, =1 otherwise.

REFERENCES

Greene, W. H. 1993. *Econometric analysis* (2nd Edition). Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Habicht, J.-P., C. G. Victora, and J. P. Vaughan. 1997. *Linking evaluation needs to design choices: a framework developed with reference to health and nutrition*. UNICEF Staff Working Papers: Evaluation and Research Series, Number EVL-97-003. New York: United Nations Children's Fund (UNICEF).

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153-161.

International Fund for Agricultural Development (IFAD). 1998. *Annual report, 1997*. Rome.

Valadez, J., and M. Bamberger. 1994. *Monitoring and evaluating social programs in developing countries*. Washington, D.C.: World Bank/IBRD.