PN-AAS-677

EVALUATING EFFECTS OF CHILD FEEDING PROGRAMS

Report of Workshop on Nutritional Surveillance in Evaluation of Food-For-Peace Programs

Cornell Nutritional Surveillance Program
Cornell University
Ithaca, New York
September 6-7, 1984

Report prepared by:
John Haaga, Larry Clark, Barry Edmonston, Jean-Pierre Habicht,
Chessa Kenrick, Kathleen Kurz, John Mason and Karen Test

April 1985

The workshop was partially funded by the USAID Bureau for Food for Peace and Voluntary Assistance, Office of Program, Policy and Evaluation.

The CNSP is supported by Cooperative Agreement # AID DSAN CA-0240 between the Office of Nutrition, Bureau for Science & Technology, USAID, and the Division of Nutritional Sciences, New York State Colleges of Human Ecology & Agriculture and Life Sciences, Cornell University, Ithaca, New York 14853.

A report of research of the Cornell University Agricultural Experiment Station.

TABLE OF CONTENTS

		Page
Executive Summary		
Background and Purpose of the Workshop		1
I.	Design Considerations for Impact Evaluations	3
I.1	Design issues	4
I.2	Types of design	6
I.3	Unit of analysis	11
I.4	Prospective and retrospective studies	13
I.5	Case-control studies	13
I.6	Feasibility study before impact evaluation	14
II.	Confounding	15
II.1	Comparability of program and comparison groups	15
II.2	Temporal comparison	17
II.3	Information bias	17
II.4	Regression to the mean	18
III.	Outcome Indicators	19
III.1	Anthropometry	19
III.2	Mortality and morbidity	22
III.3	Other outcome indicators	22
IV.	Adequacy Evaluation	23
V.	Presentation of Findings	25
VI.	Direction for Future Impact Evaluations	26
	References	28
	Appendix 1. List of Participants	30
	Appendix 2. Workshop Agenda	33

This report draws extensively on the background paper for the meeting, ref. 1, which is available on request from CNSP (as are other CNSP publications).

EXECUTIVE SUMMARY

The workshop dealt primarily with evaluations of the nutritional impact of child feeding programs operated through maternal and child health centers, though evaluation studies with other purposes and other measures of impact were also considered. The goal was to help USAID in its planning of future impact evaluations. Participants felt that impact evaluations of selected programs can produce plausible, useful, and generalizable inferences about the impact of supplementary feeding on child growth, if there is careful attention at an early stage of the program to evaluation design and data analysis. The resources and effort required for such studies are likely to be too great to meet routinely for every program; most management purposes would be met with process monitoring or adequacy evaluations rather than the more intensive impact evaluations.

Evaluation design should receive careful attention during planning and implementation of programs selected for impact evaluation. A feasibility study should consider the need for an intensive study and the likelihood of useful results. Both in design and data analysis the goal is to make comparisons between the growth of groups of children differing only in program designs participation, so far as possible. Random assignment of subjects to program and control groups is almost never feasible in practice, so quasi-experimental designs are needed. Confounding variables need to be considered explicitly. Several designs were considered; combining before-and-after, with-and-without program comparison was considered one of the strongest. Where this is impractical, alternatives include "staggered implementation", involving comparisons of new entrants (individuals or communities) with those already in the program. In all designs control for confounding factors, at both individual and community levels, is crucial.

Threats to the internal validity of studies, such as self-selection of participants, selective withdrawals, secular or seasonal trends, regression to mean, and age confounding, were discussed along with methods for dealing with them in nutritional evaluation. In most cases, there is little that can be done at the analysis stage if these problems were not considered in study design.

Of the various outcome indicators, anthropometry was considered essential for measuring nutritional impact. Comparison of group mean Z-scores

or mean percentiles or percentages were recommended, as well as the more familiar comparisons of percentages below cutoff-points defined as percents of median reference values.

Indicators of the economic value of the supplement to the family will also be useful for many evaluations. Dietary data may be needed for special analyses, but they are difficult both to collect and to analyze, and are not needed for many types of evaluation. Mortality is important for many health and nutrition evaluations, but requires indirect, often difficult estimation and often large sample sizes. Morbidity estimates were considered less commonly useful.

Adequacy evaluation - the estimation of gross outcomes rather than those strictly attributable to the program alone - was recommended for many programs. It can be used for adaptive program management and provide important information to the communities in which the program operates. Information retrieved from growth records, and, when available, follow-up information on those who have dropped out of a program, can provide insights on how well the children actually reached by the program are doing.

Specific recommendations for the presentation of impact evaluation findings and the sequence of planning large-scale studies, beginning with feasibility study and outside review, were also made.

BACKGROUND AND PURPOSE OF THE WORKSHOP

The purpose of this workshop was to discuss issues relating to evaluation of the nutritional impact of supplementary feeding programs as part of maternal and child health programs, to provide guidance to USAID's Bureau for Food for Peace and Voluntary Assistance, Office of Policy and Program Evaluation (FVA/PPE) in planning its strategy for the next generation of impact evaluations. In preparing for the workshop, CNSP staff wrote analytic summaries of selected recent evaluations (most of which were performed for USAID) and drafted background notes identifying some of the important problems of study design, data analysis and interpretation. The evaluations reviewed were of feeding programs in India, Morocco, the Philippines, Senegal, and Sri Lanka (see list in references). The background notes were intended to provide a basis for discussion, rather than an exhaustive review, and are available from the Cornell Nutritional Surveillance Program (ref. 1). This document reports the conclusions reached at the workshop and the recommendations offered by the group to FVA/PPE for its current planning. Earlier drafts were circulated to all participants, and their comments have been used in preparing this report.

The focus of the workshop was primarily on impact evaluation, with some discussion of adequacy evaluations. The terms used at the workshop for different types of evaluation were as follows:

Process evaluation involves answering questions like "How is the program doing? What is the program doing? Is enough of the right food getting to the right people at the right time?" **Adequacy evaluation** goes beyond these, to look at gross program outcomes, addressing questions like "How are the children in the program doing?" **Impact evaluation** attempts to ascertain the net outcomes, for example, that portion of an improvement or deterioration in the nutritional status of program beneficiaries that is plausibly attributable to the program's activities. It addresses the question, "What difference did the program make?"

These further questions about impact are not easy to answer, certainly not as a simple by-product of other program activities using information gathered for other purposes. The difficulties were discussed in detail at this

workshop. Participants felt there was little point in attempting impact evaluations for most projects, since the commitment of resources and the burden on program staff required to produce plausible estimates of impact would be infeasible. Impact evaluations of the type discussed at the workshop - in contrast to the more routine types of evaluation - generally require resources from outside the program; often the results are more useful in setting policy for future programs than in helping the program being evaluated. Nonetheless, all participants agreed that impact evaluations can produce reliable information, and that selected impact evaluations are needed for both AID and the PVO's to achieve their goals.

The workshop was structured around three working groups, dealing with evaluation design, confounding variables, and outcome indicators. The overlap of these topics was recognized: for example, the most important concern in evaluation design is to reduce the effects of confounding (factors other than the affect of the program accounting for differences in nutritional status between program and non-program groups). This structure has been followed as well in this report, but we wish to emphasize that these three aspects (design, confounding, choice of indicators) are interlocking. This report should be viewed as an attempt to describe in sequence what is actually a multi-faceted subject.

We should also stress that this report is not meant to stand alone as an introduction to evaluation or as a manual. Concepts and terminology are not defined in detail here. A list of sources generally used by participants is given at the end of the references. Also, the background paper for the workshop (ref. 1) should be used in conjunction with this report. The view that producing another manual is not a priority at present was shared by workshop participants. Rather, participants agreed that application of existing techniques and the accumulation of case material from careful studies in the near future are the way ahead.

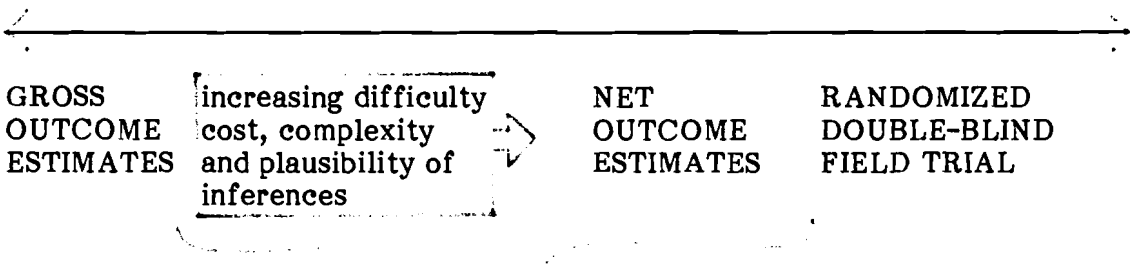
The first section of this report deals with overall questions of study design, starting from the different types of information, and degrees of plausibility, needed by policy makers, and working back to the study designs needed to produce the information. Section II deals in more detail with the

issues of confounding and ways in which they are handled at both the design and analysis stages. Recommendations concerning outcome indicators that could be used to assess nutritional impact are discussed next in Section III. The next section (IV) deals with adequacy evaluations - studies that do not attempt to ascertain net, as opposed to gross, program effect on nutrition. Section V summarizes recommendations for the presentation of future impact evaluation results, to make it easier for FVA/PPE and other analysts to assess findings and relate them to conclusions. The last section of text (VI) proposes a strategy for implementing these recommendations over the next few years. The annexes list participants and the agenda for the two-day workshop.

I. DESIGN CONSIDERATIONS FOR IMPACT EVALUATIONS

Study designs for impact evaluation can be seen as a continuum, ranging between one pole of adequacy or gross outcome evaluation, which can lead only to very tenuous guesses about the impact attributable to the program, all the way to the opposite pole, representing the classic double-blind trial with random assignment of subjects to treatment or control groups, which would lead to very solid inferences, at least about the impact on one particular population. In between, certain estimates of net outcome (or impact) can be obtained by careful design and analysis, and this was the main topic of the workshop. On this continuum, as one proceeds from left to right, the plausibility of inference, the difficulty and complexity of the study in field conditions, and its costs, are all increasing:

FIGURE 1. CONTINUUM OF EVALUATION DESIGNS



Range for feeding program impact evaluations discussed in this report

Therefore the decision about level of achievable plausibility is a compromise between the difficulty and cost. Workshop participants recommended that some evaluations should be done with a high degree of plausibility, that is, toward the right-hand side of this continuum. Only a relatively small number of these need to be done each year, but these few evaluations should have strong designs. For reasons that will be discussed below, impact evaluations of operational-scale programs will never attain the ideal of the randomized double-blind trial. However, with appropriate design and analysis, many of the threats to validity of inference that the ideal trial would eliminate entirely can be mitigated.

The evaluation's intended use must guide the decision about the level of plausibility that is needed. A high degree might be necessary to a government with severe budget constraints which wanted to be sure that the program of administering a food intervention had large enough nutritional or health benefits to justify taking over its financial support from an international agency. Funders from international agencies themselves might need almost this same level of plausibility if they are deciding whether or not to continue a program or whether to expand it. On the other hand, adequacy evaluation will suffice for many other program purposes, such as interim funding or routine management decisions.

In the remainder of this section, we discuss:

(1) general considerations of evaluation design for feeding programs;

(2) types of design that can be used for an evaluation; and

(3) suggestions for a feasibility study which should precede intensive impact evaluations.

I.1 Design issues

The crucial task in both design and analysis for impact evaluation is the control of confounding variables. A confounding variable is one that affects children's nutritional status independently of the feeding program, but is also associated with program participation. Changes in nutritional status due to the confounding variable might be mistakenly ascribed to the program. For example, children who participate in a feeding program might have better access to health care than those who do not participate, and therefore might grow better than non-participants irrespective of the

feeding program; in this example, access to health care would be a confounding variable, the effects of which on nutritional status should not be ascribed to supplementary feeding per se.

This can work in either direction, either exaggerating or understating the true impact of the program. The basic principle for evaluators is to try to compare groups of children that are as alike as possible in terms of age, family income, social status, environmental hygiene, and other conditions of life, and that differ primarily in whether and to what extent they participated in the specific program being evaluated. Inadequate control can invalidate, or call into serious question, interpretation of evaluation results. The major types of confounding are considered in section II in more detail.

Threats to the validity of inference comes from measurement bias, regression to the mean in screening programs (below, Section II), and self-selection of participants. These are issues that must be dealt with in program design. For instance measurement bias can be dealt with at the design stage by ensuring that measurers are assigned across the important groups to be compared. In analysis, measurement bias can be estimated and corrected if parts of the sample have been measured twice by different measurers, either as a post-enumeration survey or as overlapping assignments.

Misspecification of the intervention, for example, attributing results to supplementary feeding alone when feeding plus health care was the intervention, is avoided by correctly attributing results to a package of nutrition and health services or by separating out the effects of the various components of the package. The type of misspecification can best be avoided by structuring the research to address this sequence of questions: (a) Is there a nutritional impact of the total package? (b) If yes, can the parts of the package be disaggregated? and (c) If they can be, what is the nutritional impact of the rest of the package after one part has been removed, compared to the whole package? That is, if the goals of an evaluation were to estimate the impact of a package of health services including supplementary feeding, and to estimate how much of this impact is due to feeding alone, then a three-way comparison might be set up: the growth of children not in program areas vs. those with access to health services alone vs. those with access to health services and supplementary feeding. This type of research design is difficult, but can produce very

important information, otherwise unobtainable, as the Narangwal study, for example, has shown (ref. 3).

Evaluation study designs must often trade off internal and external validity. Internal validity is the degree to which the results of an evaluation accurately represent what has happened to nutritional status and its determinants in the particular population being studied. This is a precondition for any extrapolation of inferences to a larger population. External validity is the degree to which results for a sample or for one population can be extrapolated to others. If a study has high external validity, and results can be extrapolated, then policy decisions can confidently be based on these results. Evaluation studies of operational-scale programs should not be so specific to one particular area covered by the program that they can say little about the program as it operates elsewhere under normal conditions. One problem with some of the studies discussed at the workshop, often unavoidable, is that they relied on data generated by exceptionally well managed clinics; this must be acknowledged in evaluation reports as a limitation on external validity.

Data collection and analysis should focus on those ages, or malnourished groups, most likely to show impact. For pre-school children, it is likely that the greatest impact will be seen for the weaning age group (about 12-36 months); including a wider age group in analysis may only dilute the analyst's ability to detect an impact among the children most susceptible to improvement.

I.2 Types of design

Table 1 shows a typology of research designs that have been used in impact evaluations of supplementary feeding and other types of health care and nutrition interventions. These are all "quasi-experimental" designs, in that they do not call for random assignment of individuals to "treatment" or "control" groups, which is usually impossible for operational-scale programs in field settings. The reliability of inferences about program impact produced by these different types of studies depends on how well they reduce, or at least make it possible to measure, the effects of confounding variables on the differences between program participants and the children to whom they are compared. The designs are classified in two ways. The first is by the type of comparison being made: program participants vs. non-

TYPOLOGY OF COMMON IMPACT EVALUATION DESIGNS

COMPARISIONS	UNIT OF ANALYSIS	SUITABILITY FOR FEEDING PROGRAM IMPACT EVALUATIONS	COMMENTS
1. With program vs Without	A. Individual	X	Requires very careful treatment of confounding. Self-selection big problem.
	B. Community	X	Individuals in a non-program area may travel to participate, and self-selection is a problem.
2. Length of time in program or degree of participation	A. Individual	+	Self-selection and aging effects largest problems.
	B. Community	+	Problems similar to 1.B.
3. Before and after program	A. Individual	+	Aging effects, Secular trend, Pegression to mean (if screening). Need for follow up withdrawals.
	B. Community	+	Secular trend most important confounding.
4. Compare new entrants to those already in program	A. Individual	+	Same as 2A
	B. Community (Staggered implementation)	++	Potential confounding if reason for delayed implementation independently affects nutritional status.
5. Combination of with vs without program and before-and-after	A. Individual	++	Strong design, deals with aging, secular trend, etc. Still control needed for self-selection.
	B. Community	+++	Strong design, Political/ ethical difficulties in keeping some communities without program.

Legend: X Not a strong design - potential for confounding is too high, unless treatment is randomized

+ Somewhat recommended

++ Recommended

participants, long-time participants vs. those with shorter or less intensive participation, participants before and after the program, long-time participants vs. new entrants; and a combined strategy of before and after comparisons for those with and without programs. The second classification is by the units being compared: individual program participants (vs other individuals) or communities in which programs operate (vs other communities).

The first type of individual-level comparison (1A in Table 1), between participants and non-participants, measured either at one point in time by a survey, or over time by means of growth records, is usually vitiated by self-selection of participants. Self-selection is hard to avoid, or even identify. For example, children brought first, or most often, to the program, may be those with the most concerned caretakers, or those whose caretakers have sufficient time to attend to their needs, or those whose caretakers are the most adept at using public services. In these cases, the program participants would tend to have better health and nutritional status than non-participants, even non-participants living in the same environment. Alternately, if a program is locally considered a curative health intervention, or if there is a social stigma attached to program participation, those who receive supplementary feeding may tend to be the children with acute or chronic illness, or the members of social outcast families or broken homes. Comparing their health and growth with those of the more fortunate non-participants would tend to obscure a beneficial effect, if the program had one. If the relevant social and economic variables were easy to measure accurately, and their independent effects on nutritional status could be estimated and entirely controlled for in analysis, then self-selection would not be a threat to valid inference about program impact, but neither condition ever holds. Evaluators should try to measure social and economic background variables in any case, partly because of their potential for confounding and partly because they are intrinsically interesting, showing how well a program is reaching its intended beneficiaries. Simple comparisons of participants vs. non-participants, however, are very likely to be misleading.

Comparisons between program and non-program <u>communities</u> (1B in Table 1) are subject to similar effects, though perhaps to a more manageable degree. Communities may "select themselves" for a program by having more active and efficacious leaders, or just by being more accessible geographically; in either case, their health care,

sanitation, water supply, access to any income-enhancing public services, etc., are also likely to be more favorable than those in non-program communities. All this depends, of course, on how communities were selected for the program, which is a major concern for evaluators. Also, it is sometimes hard to sort out program from non-program communities; individuals may travel from one place to another to participate (or if supplementary feeding is really valuable, as in a food crisis, may even relocate their residence.) For example, a study in Senegal found large numbers of participant children in non-program villages (see citation and discussion in reference 1.) Occasionally, it is difficult to define participation, if, for example, children not participating directly share food with others in their household who receive rations.

Similar disadvantages attend studies of the second type of design (2A and 2B in Table 1), comparing the longer or more intensely participating individuals and communities with those which have participated for shorter times, or to a lesser degree, or less frequently, or with long interruptions. Again, the problem is that length, frequency, and continuity of participation are likely to be related to other individual and community characteristics affecting health and nutritional status independently of the feeding program. However, this design may be stronger than the simple with vs. without program design, because a finding of a "dose-response" relationship between program participation (as a multi-level or continuous variable) and nutritional status would often be more convincing, at least as presumptive evidence of impact, than the simple two-way comparison. As with designs 1A and 1B, inferences about program impact would still be weak, and would require a very careful attempt to measure and estimate the effects of confounding social and economic variables.

"Aging" or maturation effects are a problem for any designs which call for comparing children of different ages. In type 2A, for example, the children who have been in the program longest may simply be older than those who have been in a short time, and in a before- and after- comparison of individuals (type 3A in Table 1), the "after" measurement is necessarily taken when the child is older than when the "before" measurement was taken. The problem is that the low scores on the various anthropometric indicators are more common at some ages than others; children may improve or deteriorate on these scores simply because they have passed from one age group to an older one, not because of a positive or negative impact of the feeding

program. Care must be taken in analysis to compare groups with similar age composition (e.g., all children aged 6 to 60 months in a community before and after the program, as in type 3B) or to compare children's weights or heights to what would be expected locally, given their age at time of measurement. These problems are discussed further in sections II.1 and III.1, below.

Before-and-after comparisons are also difficult when there is screening into a program (e.g. children are eligible only if they have low weight-for-age) or graduation from a program (e.g., children are no longer eligible if they have gained weight satisfactorily for several months). When there is screening, regression to the mean becomes a problem (see section II.4 below), and a simple before-after design would be especially misleading. Graduation may similarly invalidate simple before-after comparisons unless there is a way of obtaining some information on former participants (see section II.1).

Both individual- and community-level before-and-after comparisons are subject to the problem of secular trend, i.e., non-program related changes affecting health and nutrition. There might be steady improvement (or deterioration) in standards of living, hygiene, and health care, or even rapid changes due to new employment opportunities in the community, or natural disasters, or wars. All would make the before-and-after comparison an unfair test of program impact, unless their effects were small enough and measurable. The longer the period between first and last measurements, the more secular trends are likely to account for observed changes. This effect can be estimated partially by comparing mortality rates, growth data, immunization coverage, etc., in study areas with any available provincial or national level data covering the same periods. Incorporating comparison communities into the study design (as in types 4B and 5B) helps estimate the secular trend, especially if conditions in non-program communities (5B) or new program communities (4B) are continuously monitored.

A combination of the before-and-after and the with- and without-program designs (types 5A and 5B in Table 1) guards against many of the threats to invalid inference discussed above. Two groups of individuals or communities are measured in a baseline survey, those that will and those that will not have a chance to participate in the feeding program. These groups are chosen to be as similar as possible in all respects other than program participation. Unless these other factors change

unexpectedly in one but not the other group, differences in nutritional status that emerge during the life of the program can reliably be attributed to program activities.

A major difficulty with designs of type 5 is that it is often infeasible, administratively and politically, to keep a group of individuals or communities unserved by the program for the length of time needed to observe differences in children's nutritional status due to the program. It is also a challenge to researchers and managers to devise an ethical study, often requiring provision of health services other than those included in the program being evaluated to the non-program communities. If the non-program communites are near the program communities, then they may not be sufficiently separated for valid comparison. Besides the possibility that people in non-program communities benefit directly from the program, there is a possibility of indirect benefits: the spread of health and nutrition knowledge, for example. If a program causes a sufficient increase in the food supply of one community to have a nutritional impact, then it will also have an impact on the food economy of the surrounding communities, as supplies increase and prices are lowered, either through a replacement effect or actual re-sale of program commodities. If these problems are avoided by choosing comparison groups far apart geographically, then the groups may differ on too many other factors, leading to a shortage of "degrees of freedom" for the analysts.

Such practical considerations often force a sort of compromise, in the form of staggered implementation of a program (type 4 in Table 1). Rather than attempt to keep some communities unserved for years, socio-economic conditions, child health and nutritional status are measured in each community as the program expands to it, then re-measured in later years. This allows a type of with- vs. without-program comparison (the communities just entering the program are the "withouts") and also a dose-response comparison like that of design type 2, comparing communities served longer with those in the first or second year of a program. Secular trends can be estimated by the observed changes over time in the measurement in new entrants. The major problem is posed if the communities first enlisted differ systematically from those later enlisted in ways likely to affect nutritional status, e.g., accessibility. In all of the designs discussed here, it is important to measure potential confounding variables and include them in analysis, not just assume that the study design somehow takes care of them.

If there is more than one component being expanded at the same time, it may be possible to vary the sequence in which they are introduced. This would produce some element of variation needed for measuring separate impacts, and also avoid some of the administrative, political, and ethical difficulties. (This procedure also produces information needed for cost studies - again, a model is the Narangwal study, reference 3).

I.3 Unit of Analysis

An important consideration is the unit of analysis of the evaluation study, the community or the individual participant. (The typology of designs, Table 1, divides types of analysis into these two larger categories). The chief difficulty with the individual-based studies is that self-selection is hard to avoid, even to identify. For example, the children who happen to be brought first or most often to the program may be those with the best caretakers, who would be better off anyway, even without the program. Or they may be only those with acute illness, if the program is considered locally to be a medical intervention. In either case, measurements of nutritional status and of potential confounding variables at entry may be needed to detect lack of comparability. Also, in many studies, it may be difficult to assign individuals correctly to the groups being compared; children not counted as participants may actually be benefitting from a supplementary feeding program if their family members or friends are participants.

If there is screening into a program or graduation from it (i.e. selection of participants by those operating the program, in addition to self-selection by participants' families), this provides an additional complication for individual-based studies and a further reason for community-based studies. Children in the program are then, by definition, a select group, not representative of the whole population of the community and not comparable to non-participants for purposes of estimating impact.

Comparing the situation of all children in program communities with that of all children in non-program communities obviates these difficulties. Many communities may have to be included in the evaluation, however, which becomes a practical problem. This is because communities differ in many other ways affecting child health and nutritional status, besides program participation. The program and non-program

groups must either be exactly comparable in terms of existing health services, diets, housing quality, water supply, sanitation, family incomes, etc., or else there should be enough variation in these within each group to sort out the effects of the program from those of the other factors.

If the relevant characteristics of communities were all measurable and known before program implementation, then the number of communities required to assure comparibility in the analysis is relatively small. But in discussions about these topics at Cornell subsequent to the workshop, some participants felt, based on the experience of supplementation field trials with small numbers of replicates, that matching in advance is so imprecise that larger numbers (say, tens of communities in each group) are needed in the future. Places thought to be comparable in terms of health services or socio-economic conditions when a study is planned often turn out not to be, or, if initially comparable, subsequent events like the building of a new road, military base, or whatever, may wreck the balanced design. (By way of illustration, recent calculations for an evaluation design in Asia indicated that some 20 or so replicate villages were needed in <u>each</u> group to give the required degrees of freedom for analysis).

Analysis of data from program and non-program communities is likely to include child-level characteristics (age, sex, parity, perhaps birthweight or weight at entry to a program) family-level characteristics (e.g. income, housing quality, household water supply and sanitation) and community-level characteristics (notably program availability). Even if it is only the effect of the program that investigators are interested in, the other variables affecting nutritional status should be included, partly to remove bias in the estimate of the effect of primary interest and partly to test for interactions or threshhold effects (see ref. 1, and other references listed, for definitions and technical discussions).

The "effective sample size" for such analyses is somewhere between the number of individuals and the number of communities — exactly where depends on the relative size of inter-community variation in nutritional status and intra-community variation. The effective sample size for these comparisons may not be increased much, after a certain number of children, by measuring more children in each community; instead, the cost-effective way to increase effective sample size is to increase the number of communities. The precise formulas for determining sample sizes can be found in many

statistics texts (e.g., refs. 5 and 6). As a rule of thumb, in the national nutrition surveys sponsored by the USAID Office of Nutrition it was not generally found worthwhile to measure more than thirty children per community; and indeed that number is high in relation to many household surveys where typically 10-20 households per cluster are used.

I.4 Prospective and Retrospective Studies

Evaluation studies can also be classified by whether they employ prospective, retrospective, or a mixed strategy for data collection. Prospective studies are those in which data collection for purposes of evaluation begins before the program itself begins in the areas to be studied; the study is carried on in "real time". Retrospective designs for data collection call for retrieval of data some time after the event or measurement, e.g., by collecting information from growth cards or by eliciting mothers' recall of child deaths or morbidity. These data collection strategies can sometimes be mixed, for example, by using prospective methods for some variables, or for sub-samples, and retrospective methods for other variables or larger samples of children.

A retrospective data collection strategy is sometimes appropriate, but it should not be confused with belated planning of an evaluation study. Whatever the type of data collection, if an impact evaluation is to produce plausible results, the evaluation design should be part of initial program planning. The practice, occasionally followed by some international agencies, of calling in outside evaluators for a rapid appraisal after a program has been operating several years with haphazard data collection and no planning for eventual evaluation, can lead only to impressionistic reports, not to valid inference about impact.

I.5 Case-Control Studies

Another study design, the case-control study, was also discussed at the workshop, but opinions differed on its likely usefulness for impact evaluations of feeding programs. In this design, subjects with a certain condition are sought, either by screening or by a search of records, then matched with other subjects on the basis of factors not being studied (e.g., age and sex). The two groups, cases and controls, are then compared for the presence of one or more "risk factors", often measured by

retrospective data. If there are no biases introduced by the process of finding cases, and if variables are measured accurately, this can produce estimates of the relative risk of the condition associated with the risk factor, answering, for example, the question "How much does past or current participation in a feeding program affect the odds that a child is malnourished?" This procedure does not produce any way of estimating the prevalence of the condition in the community. Nor does it remove some of the other important problems associated with other individual-based comparisons, caused by self-selection of program participants on screening and graduation criteria. In subsequent discussions at Cornell, it was suggested that this design might be most appropriate when the condition being studied is relatively rare in a community (e.g., xerophthalmia or child mortality), since the economic advantages of case-finding as opposed to household surveys would then be greatest. But for feeding programs assessed by measuring child growth, case-control designs do not generally offer an advantage.

I.4 Feasibility Study before Impact Evaluation

Given the importance of obtaining plausible and accurate estimates of impact for selected programs, and the difficulties outlined in previous sections, FVA/PPE should consider formalizing the procedure of a detailed feasibility study before resources are committed to a major evaluation. (A similar practice was followed for at least one of the case study evaluations discussed at the workshop.) This feasibility study should first propose specific audiences for the evaluation, set out the questions to be answered, with an idea of their priority for policy makers, and thus make recommendations on the need for a particular study and the level of effort. To decide initially whether impact evaluation of a certain program is worthwhile, a feasibility study should include a preliminary process evaluation. If preliminary evaluation of program delivery indicates problems, then it is unlikely to be worth trying to evaluate impact. In this case evaluation should concentrate on management and logistic issues.

A feasibility study should consider possible evaluation designs, the data needed to control potential confounding in design and in analysis, including the selection criteria for villages or individuals. If there is not enough data on variables which are potential confounders, it may not be possible to attribute results to the intervention, making the program a poor candidate for intensive impact evaluation.

II. CONFOUNDING

The major topics addressed under the general heading of confounding were (1) comparability of program and comparison groups, (2) temporal comparisons, (3) information bias, and (4) regression to the mean (see Table 2). Each potential confounder was discussed with regard to its frequency of occurrence, importance, ease of detection, possible remedies in program design or analysis, and severity if not remedied; brief conclusions on these points are given in the columns of Table 2.

II.1 Comparability of Program and Comparison Groups

Evidence should be presented in an evaluation report to demonstrate the comparability of program and comparison groups. If groups are alike with regard to characteristics such as presence or absence of potable water, health services, etc., then differences between them in nutritional status may be attributed reliably to the feeding program. To address problems of non-comparability between groups due to withdrawals, self-selection, secular trends, and other common sources of confounding, data on specific confounding variables are needed. These variables relate to:

(a) the community (e.g. potable water, distance to clinic)

(b) the family (e.g. socio-economic status, maternal education)

(c) the child (e.g. sex, age, birth order, immunization status)

(d) temporal conditions (e.g. season, year).

Many of these variables are always important and can be specified for collection in advance. Others are country-, or situation-specific and are only needed in certain instances. Specific causes of non-comparability are summarized below.

Self-selection bias. Individuals who participate in feeding programs are likely to differ from non-participants in ways that may independently affect the outcome under investigation (see above, section I.2). Designing a program to minimize the effects of self-selection does not guarantee that it will not occur. In the evaluation, therefore, bias introduced by self-selection should be assumed present and data produced to demonstrate otherwise. (An example of how such an analysis might be presented is illustrated on page III.13 of the background document, ref. 1).

Withdrawals. Three reasons for program withdrawal should be considered:

TABLE 2

SOURCES OF CONFOUNDING IN FEEDING PROGRAM IMPACT EVALUATIONS

SOURCE OF CONFOUNDING	RELATIVE FREQUENCY	IMPORTANCE	EASE OF DETECTION	REMEDY: DESIGN	REMEDY: ANALYSIS*	SEVERITY OF CONFOUNDING IF NOT REMEDIED
A. **Comparability of Groups Compared**						
1. Non-comparability of program and non-program groups	almost always	variable	feasible with right data	can be handled	difficult	varies, can be be important
2. Self-selection	always in individuals often across communities	important	detect using independent data	difficult	difficult	can be strong
3. Withdrawals**	graduation common with screening	very important	difficult	can be handled	difficult	difficult
4. Age Reporting	often a problem	important if screening	requires demographic skill	difficult	none	moderate
5. Ageing	always	high	not difficult	can be handled	difficult	moderate
6. Controlling variables	always	very important	—	modest cost	unable to handle	moderately easy
B. **Temporal Comparisons**						
7. Secular trends	sometimes, but variable	low, exception long-term	usually feasible	can be handled	feasible	moderate
8. Seasonality	usually a concern		detect by comparing measurements in same season	can be handled	impossible, unless, comparison group group available	can be expensive
C. **Misclassification**						
9. Deliberate	occasional if screening used	important if screening used	usually easy	difficult	no solution	varies***
10. Inadvertent	common if cutoffs used	moderate	difficult	yes	difficult	usually important
D. **Regression to Mean if no control group**	always when screening	very important when screening	varies, can be difficult	expensive	no solution	can be lethal

* Extent to which analysis can help, if confounding was not remedied by design.

** Includes graduation, drop-outs, and mortality. See text for full description.

*** Severity of effect depends on the variable misclassified.

(a) Dropouts - Program participants who drop out may tend to be those who find the program ineffective or irrelevant; on the other hand, they may tend to be those who have improved in health and growth and no longer feel they need the program. Families who drop out may tend to be those for whom continued participation is costly, in terms of transport time and expense or the opportunity cost of mother's time. Or families may drop-out for a specific, non-program-related reason (e.g. migration).

(b) Graduates - Each program has a different policy or criterion for "graduation". For example, children may graduate because their nutritional status has improved, not improved, or stayed the same but exceeded a time limit set for program participation. Depending on the reason for graduation, the direction of bias will differ and must be taken into account in an evaluation. If graduation is a sign that a child is doing well, then for some purposes it can be considered a program outcome in its own right. As with voluntary dropouts, having information on the subsequent health and growth of former participants is valuable.

(c) Mortality - Differences in mortality rates between participant and controls will confound comparisons if not accounted for. If a variation in mortality between groups does exist, it is in its own right an important outcome indicator and should be analyzed using life table methods. (see Section III below).

Evaluation studies should present data on the distribution of withdrawals by reason for withdrawal (as illustrated on page II.17 of the background document, ref. 1). To assess the comparability of withdrawals and remaining participants in the analyses, the last available anthropometric measurement for withdrawals should be compared to those of children remaining in the program. Differences indicate the likely direction of bias.

Aging or maturation effects. The risk of malnutrition may differ at different ages. This difference in risk is a problem when groups of children with different age distributions are compared cross-sectionally or when a cohort of children is followed through time and the effect of a program is measured as the difference in nutritional status at entry compared to a later period. In the first case, little can be done in the analysis except to assess the direction of the potential bias introduced. The latter problem can be addressed by appropriate analysis; by modeling the growth curve of the cohort and analyzing the residuals (see section III.1).

II.2 Temporal Comparisons

Secular Trends. This refers to general improvement or deterioration in the nutritional status of all children in a population occurring over time and poses a particular problem in evaluation studies when comparison groups are not available (e.g. types 1,2 and 3 in Table 1.) To ensure that either an improving or declining secular trend is not confounding evaluation results, data on nutritional status, infant or child mortality rates etc. from national surveys or other independent sources should be presented (see suggested format on page IV.16 in the background document, ref. 1).

Seasonality. Where they exist, seasonal effects are often strong and can seriously bias evaluations in which there are no comparison groups. Simple before- and -after comparisons are particularly liable to bias due to seasonal effect, if measurements are made at different times of the year. When seasonality is a problem, the indicator height-for-age is less sensitive to seasonal effects than is weight-for-age (see Section III). Young children are also more likely to be affected than older children.

II.3 Information bias

Information bias is the result of having more accurate information about one of the groups being compared than about the others. It can be minimized by training, calibration of measuring instruments, and by keeping enumerators or health personnel motivated. The more common problems discussed at the workshop are summarized below.

Misclassification bias. When criteria are used to determine eligibility for a program, as is often the case, misclassification is a potentially serious problem. Misclassification can be either inadvertent (random) or deliberate (so as to include children not meeting the entry criteria). When misclassification is random, comparing means instead of prevalences will reduce the magnitudes of error in the analysis. When deliberate misclassification occurs, the problem is more serious, and may result in overestimation of program effects. If the original data on weights, heights and ages are available they can be examined to see if clumping (e.g. large numbers of children falling just below the cutoff) is occurring.

Age. More care is often taken in ascertaining the ages of participants than of controls, leading to potential bias in age reporting. As a general technique, the date of birth should be asked, as well as the child's age, and these should be checked against each other. Some studies have found that Asian mothers tend to report smaller children as younger than they actually are (ref. 4); this practice leads to underestimates of the prevalence of malnutrition. To check if age misreporting is a problem, the age distribution of the sample can be compared to the known age pattern (or model life tables) for that area or country. Where age misreporting is common weight-for-height may be a more useful indicator. Alternatively, measures of growth velocity, which tend to be linear over a certain range, could be used. Data on a second indicator of nutritional status, such as mid-upper arm circumference, should also be collected whenever possible and used as a confirming indicator.

II.4 Regression to the Mean.

Regression to the mean refers to the tendency of individuals selected on the basis of especially low or high scores on an initial measurement to have scores nearer the population mean on a second measurement. In this context, the individuals screened into a feeding program on the basis of a low initial anthropometric score will tend to have higher scores the next time they are checked. The first score may have been unusual because of random errors in measurement and random fluctuations in weight or height themselves, but these will tend to be more 'normal' with subsequent measurement. "Regression" in this sense can be mistaken for a time program effect if one follows only the

individuals initially selected on the basis of a low score is always a potential problem when nutritional screening criteria are used for program entry. There are two ways of dealing with regression to the mean: (1) have a control group with comparable initial values on the screening indicator or (2) remeasure the entire population which was originally screened and compare the two population means. Unless regression to the mean is addressed, program effects cannot be inferred.

III. OUTCOME INDICATORS

Several indicators were discusses as measures of program impact through changes in nutritional status. The emphasis was on anthropometric indicators; mortality, morbidity, food consumption, economic, education and biochemical indicators were also discussed. The choice among them depends on the purpose of the evaluation and the goals of the program itself, as well as on technical considerations of feasibility of collection, responsiveness, etc. Desired characteristics of indicators included accuracy, ease and low cost of collection, usefulness for program management and the responsiveness of the indicator in question to likely changes in nutrition status. Table 3 summarizes many of the points discussed under these headings.

III.1 Anthropometry

The choice of an anthropometric indicator partly depends on the nutritional status of the population before intervention. For example, if mean weight-for-height (WH) in the study sample is low at baseline, then an increase in WH could be expected as a result of the feeding program. If WH is normal, however, the feeding program is unlikely to lead to significant increases.

Height is sensitive to long term impact, and is less sensitive to seasonality than weight; however, in some settings it may be more difficult to measure than weight. For this reason, height should probably not be included in routine data collection in a clinic setting, but may be part of special surveys done periodically to assess impact. If height is used, analysts need to ascertain whether length or stature measures were taken and use appropriate reference standards. Routine growth monitoring in clinics typically uses weight charts to assess nutritional status of individual children; often this will be the only indicator available for evaluation.

TABLE 3 - SUMMARY OF OUTCOME INDICATORS FOR EVALUATION STUDIES

INDICATOR	RESPONSIVENESS*	COLLECTION	OTHER ISSUES
Height-for-Age	Measures chronic malnutrition thus less sensitive to acute events such as seasonality. Responds to supplementation in field trials.	Costly to collect height data, requires trained personnel.	Means should be compared as well as proportions below cutoff points. Percents of median and Z-scores both require age standardization.
Weight-for-Age	Chronic and acute malnutrition	Weight easy to collect,	
Weight-for-Height	Measures acute malnutrition. May not be low to begin with, thus unlikely to respond to supplementation.	Weight easy to collect, Height difficult to collect.	
Arm Circumference	Less reponsive than HA or WA.	Similar to height data	Useful for verification of other anthropometric measures.
Mortality	Responsive to health care components and large improvements in nutrition	Requires recall data, often subject to biases.	Difficult to detect short-term trends without large samples. Requires demographic skill for indirect estimation of rates.
Morbidity	Not good measure of impact of feeding programs; food supplement can affect severity of disease not frequency.	Reporting often unreliable.	Maybe useful in own right.
Food consumption	Should respond if nutrition education effective.	Very expensive and time consuming to collect.	Recommended if specific problem foods are implicated, as in Vitamin A deficiency.
Breast Feeding	Should respond, except if prevalence already high without program.	Easy to collect.	Often useful for program purposes, education.
Biochemical and Clinical	Can be accurate measures, but this outweighed by feasibility	Much too expensive. Requires trained personnel to collect and diagnose.	
Economic	If supplement large, should have effect on family income and expenditures	Complete expenditures and diets difficult to collect but useful to estimate value of supp. as % of family food exp.	
Birthweight	Good, if maternal supp. part of program.	Difficult to obtain accurate measures.	

* Responsiveness partially depends upon the nutritional status of the population to begin with e.g. if mean weight-for-height is low in a population then WH would be responsive to change.

Values of weight-for-age (WA), height-for-age (HA), or weight-for-height (WH), can be expressed either as percentages of a standard value (the median of a reference population) or as centiles or Z-scores, showing where the measured child would "fit in" to the distribution of the reference population. Percentages of median are easy to compute and have the advantage that many public health workers are familiar with them, but their big disadvantage is that the variance (of both the reference population and population of children in developing countries) increases with age: a 12-month-old child whose height is 90% of the reference median is more unusual, and more severely malnourished, than a 60-month old child whose height is 90% of the reference median for that age. Expressing heights and weights as centiles or Z-scores does account for the increasing variance with age even among healthy children, but may introduce some other technical difficulties, due to the "splicing" of reference populations in the standards recommended by WHO for international use. If the groups being compared have roughly similar age distribution, however, the choice of indicator is unlikely to affect inferences.

It is important to note, that use of Z-score or centiles does not by itself "age-standardize" a comparison; the Z-scores are standardized only with respect to the reference population. For example, a program group might decline from mean HA of -1.0 (Z-score) at age 12 months to -1.5 at 48 months; this could nonetheless be evidence of a beneficial effect if the usual pattern in the population was a decline to -2.0 at age 48 months.

An alternate method of analysis discussed at the workshop, one that would not require choice of a reference population, would be to assign a score to each child equal to a residual from a regression of weights or heights (in kilograms or centimeters) on age and sex. This procedure also has the advantage of avoiding the needlessly contentious issue of choice of a reference population; in effect, the analyst is generating a local reference standard based on the entire sample of heights or weights available and comparing sub-samples (like children from program communities vs. children from non-program communities). The procedure does have the important disadvantage, besides unfamiliarity, of making comparisons of malnutrition prevalence with other populations very difficult. Perhaps evaluators could use "internal standards" in analysis, but present results in the more familiar terms of international standards, and thus be able to assure readers that the results, one group of

children being heavier or taller than another, do not depend on the choice of anthropometric standards. More work on this method is needed before it can be recommended as standard practice for evaluators.

Tests of group differences in the mean of an anthropometric indicator have more statistical power than tests of differences in proportions below a cutoff point; the latter are nonetheless often more meaningful for policy decisions like resource allocation. Analyses can be performed with both means and proportions below customary cutoff points, and presented in whatever way is useful for the policy audience.

Velocity data may be preferred to measures of attained growth, because they are more responsive to supplementation. Caution must be exercised in drawing conclusions from growth velocity data, however, because measurement error is compounded by the process of subtraction. The coefficient of variation of velocity measurements may be much greater than that of attained growth measurements. If measurements are taken every month, for example, then purely random measurement error, or chance fluctuation in actual weights, will be large relative to the apparent difference between successive measurements. Such error will be a smaller component of variance in velocity measurements, and thus less likely to mask true differences between groups, if measurements are separated by several months.

Age estimation

Because of the difficulties of ascertaining exact ages of children in countries where birth registration is incomplete or inaccurate, age-independent anthropometric indicators (WH or arm circumference) are attractive. Unfortunately, they are complements to, rather than substitutes for, the age-dependent measures (HA or WA), since they measure wasting rather than cumulative deficits in linear growth. Populations have often been found to be chronically malnourished, as evidenced by low HA, though near the reference values for WH. Accordingly, evaluators will nearly always have to deal with the problem of age estimation.

Random error in reported ages biases the estimates of proportions below cutoff points of anthropometric indicators very significantly, but affects

estimates of mean values to a much lesser extent and without bias; this is a further reason for making comparisons of means as well as proportions below cutoff points.

III.2 Mortality and Morbidity

Reliable estimates of mortality are feasible in some studies. They typically require large sample sizes, demographically sound estimation procedures, and careful control of confounding (e.g. SES, age, health care utilization). If a feeding program is well integrated with local health services, then following up the mortality experience of program dropouts should be possible. The workshop recommended that demographic techniques for estimating mortality differentials be considered in future feasibility studies for intensive evaluations.

Morbidity is not usually a good indicator for evaluation of feeding programs, because evidence suggests that food supplements do not affect the frequency of disease, only the severity. Morbidity reporting also is affected greatly by access to and utilization of health services. Morbidity information should be collected if other program components are concerned with it, or if it is a potential confounder.

III.3 Other Outcome Indicators

When nutrition education is an important part of the intervention, changes in behavior are often feasible indicators of program impact. Biochemical and clinical indicators are not recommended in the context of program evaluations discussed in this document because of the expense and expertise needed to collect and measure them, though they are useful and feasible in other contexts. Economic indicators (e.g family food expenditure) could be useful, but do not constitute a proxy for nutritional impact. Finally, health care utilization is useful, but to analyse properly one must have a control group composed of a clinic without a feeding center. In addition, it is often difficult to separate the effect of health care use versus food supplement if the two are correlated with one another, and this issue should be addressed in all reports.

Food Consumption and Breast Feeding

Quantifying nutrient intake of an individual is very difficult and expensive. It is recommended therefore that dietary data are worthwhile to collect only in the following circumstances; when enough resources and trained personnel are available, when policy dictates that it be collected (e.g. for ration estimates), or when a specific nutrient problem can be evaluated by simplified data collection, e.g. Vitamin A deficiency. Information on breast feeding should be collected routinely because it is easy to obtain and may be asked anyway as part of an education component of a program. Care should be taken when associating breastfeeding with nutritional status; SES and age of child are frequently confounding variables.

IV. ADEQUACY EVALUATION

Though the workshop dealt primarily with issues surrounding impact evaluations, participants discussed adequacy evaluations of feeding programs as well, in working groups and plenary sessions. Questions considered included: How frequent or extensive should such evaluations be?; What timing of evaluation in the project cycle makes sense?; To what extent can adequacy evaluation be handled within a program, by its own staff and using only routinely generated data?

Discussions began by maintaining a distinction between process evaluation - the monitoring of a program's actual delivery of outputs as intended to intended beneficiaries - which should be a part of every program, and adequacy evaluation, which includes process evaluation but goes beyond it to include monitoring outcomes, in terms of health and nutritional status of program participants. A more inclusive list of information that conscientious managers ought to seek emerged during discussions: in addition to information on outputs and gross outcomes, important topics are coverage (the percentage of the malnourished in program communities participating in the program), reasons for non-participation of intended beneficiaries, and reasons for dropouts and some follow-up information on them. Evaluation in this sense starts to merge with good management; "monitoring and evaluation" in practice becomes something like what is called in other contexts "management information systems" (MIS). Participants felt that MIS should be part of nearly all programs, but that the temptation in setting up such systems is to focus too much on routine data

collection and the filling out of forms for their own sake rather than answering the important questions.

Adequacy evaluation would need to be included in program design from the start - it is part of a good program - rather than added as an afterthought. This would not, however, preclude more intensive efforts, involving evaluators not connected with program operations, to deal with these questions in depth and with an independent outlook. Such an intensive study would need to be preceded by a feasibility study (much as the workshop recommended feasibility study before intensive impact evaluation). The purpose of this would be both to do some of the preliminary process evaluation - if the program didn't happen at all as planned, or if there is no information available, it would not make much sense to examine outcomes or coverage in depth. It would also ascertain the interest of those connected with the program, either its implementors or intended beneficiaries, in the information that might be produced. There are both costs and benefits to close involvement of "insiders" in the work of an "outside" evaluation team; but most workshop participants felt that the advantages of relevance and ultimately local use of the findings were great enough that those connected with the program should usually be closely involved. This would not just be a matter of handing a completed report to program staff on participants. Rather, their expertise and insights should be used from initial stages of analysis onwards.

Where flexibility in program implementation is allowed (intentionally or unintentionally at higher levels), and where managers have been paying attention to formal or informal MIS's, programs will have been changed, even within a few years, from what was originally set out. Evaluations of feeding programs and other components of MCH clinic work are not necessarily like field trials for new therapies; the "treatment" itself changes adaptively as the situation in the community changes. This may be untidy for evaluation studies but important for workable and relevant programs.

Growth card data can be used to assess the health and nutritional status of children in the program. Such data are often obtainable from routinely generated records; evaluators should capitalize on opportunities to link data from different sources for a longitudinal record of health and nutrition trends in the program population. Some programs have required too frequent collation and reporting of clinic data. The intervention may call for growth monitoring

of each child and discussion of his or her progress with the child's caretaker, but not every weighing of every child needs to be reported to higher administrative levels. Sampling among children, or checking only once per season, may reduce the reporting burden on clinics and the information overload on managers. In assessing growth data, evaluators and managers need to compare patterns both to those of children growing normally (e.g., in a reference population) and to a typical local pattern for non-program areas or children.

While the emphasis in the discussions of impact evaluation was on anthropometric indicators of outcome, for adequacy evaluation, other, intermediate indicators were also considered important. For example, the perceived benefit of the food supplement (its value as an income transfer) and the effect of program activities on women's time use could be important. Again, the important initial question to resolve is the purpose of the evaluation.

When adequacy evaluations produce negative or equivocal results, some of the questions that must be asked are:

1. What are the program's unintended or unforeseen consequences? Is the program appropriate for this population?
2. Were the indicators used to measure or approximate outcomes appropriate ones?
3. Is the program appropriately targetted to those who would actually benefit?

V. PRESENTATION OF FINDINGS

Two main recommendations were made on the presentation of results: (1) an advisory board to FVA/PPE should be formed to evaluate the findings of impact evaluations. This board should review the design and feasibility of the next generation of evaluation studies and the findings. Its members should represent diverse professional backgrounds. (2) Three separate reports should be written: (a) a feasibility report, (b) a summary report, and (c) a scientific report.

The feasibility report should contain specific plans for the impact evaluation, including precise objectives, the outcome indicators and how they will be measured, data sources for confounding variables, sampling plan, etc. The summary report, envisioned as about ten pages in length, should contain: a brief description of methods, the main conclusions of the evaluation, with

illustrative tables and figures; and sufficient discussion of issues of confounding, internal and external validity, that those connected with the program can judge the strength of various conclusions. The summary report would thus be an expanded version of the conventional Executive Summary. In most current evaluation reports, it was felt, the executive summaries are for either audience, and lack the necessary references to sections of the detailed report where assertions are backed up. The audience for impact evaluation reports is inevitably a diverse one, but perhaps the formats suggested at the workshop will prove useful in meeting their needs. This summary report should contain ample references to appropriate sections of the scientific report, which should contain the detailed descriptions of the population studied, the sampling, measurement, and analytic methods, marginal distributions and summary statistics for all quantitative variables, parameter estimates and standard errors for any multivariate models, and similar information.

The workshop recommended separation of the summary report and the scientific report because many participants felt that previous evaluation reports have tended to fall between two stools, being too detailed and technical for policy makers to see clearly the important conclusions and their implications, but too cursory for technical advisors and other researchers really to understand what was done and comment usefully on statistical issues or replicate the analyses. Conventional Executive Summaries are often too short

VI. DIRECTION FOR FUTURE IMPACT EVALUATIONS

The workshop's recommendations for future evaluation of supplementary feeding programs are based on the following reasoning:

(1) Impact evaluation can find quantitative improvement in the nutritional status of program participants when it has occurred. There may be as much danger of incorrectly concluding that a program has had no nutritional impact, as of falsely inferring impact when there was none in reality, if studies have weak design and analysis.

(2) Evaluations produce results at various levels of plausibility and results from impact evaluation can attain a high level, if studies are carefully designed and receive a high priority for funding.

(3) Evaluators of supplementary feeding programs do not need more guidelines. Their work might be aided, however, from a compilation of

evaluation case studies from which they could consider different types of design, different types of analysis and presentation of results, and different interpretation of results.

Based on these premises, the following recommendations were made. First, USAID and other organizations who commission program evaluations need to clarify the idea of evaluation, particularly impact evaluation. They also need to sensitize groups of people to the usefulness and importance of impact evaluation. These include top-level government officials, for whom impact evaluation can guide national policy; funding agencies, for which it can guide organization policy; and Private Voluntary Organizations (PVO's), for which it guides program selection, design, and management.

Second, impact evaluations should aim for a high level of plausibility, receiving a high priority for funding, and only a few should be commissioned each year. The procedure for such evaluation should include, in the following order, (1) a feasibility study; (2) a design for the impact evaluation based on results from the feasibility study; (3) the survey and/or retrieval of relevant data; (4) data analysis; and (5) report of 1-4 and their interpretation. USAID might use a technical advisory committee or consultative group, as indicated in Section V of this report.

Third, in addition to impact evaluation, management information systems (MIS) for process and adequacy evaluation should be emphasized. Whereas impact evaluation might be commissioned for only a few programs each year, MIS evaluation should be an integral part of most supplementary feeding programs. USAID and other organizations that commission evaluation of such projects need to clarify the idea of MIS evaluation, to sponsor development and dissemination of good case studies, and to sensitize government officials, funding agencies, and PVO's to its usefulness and importance.

REFERENCES (CITED IN TEXT)

1. K. Kurz, J. Haaga, C. Kenrick, J. Mason, and K. Test, "Methods Used for Impact Evaluation of Child Feeding Programs: Background Notes for Workshop, 6-7 September, 1984", Working Paper No. 32, Cornell Nutritional Surveillance Program, Cornell Unviersity: Ithaca, NY, August 1984.

2. J. Mason and J-P. Habicht, "Stages in the Evaluation of Ongoing Programmes" in D. Sahn, R. Lockwood, and N. Scrimshaw, eds., Methods for the Evaluation of the Impact of Food and Nutrition Programmes, (Food and Nutrition Bulletin supplement 8). The United Nations University: Tokyo, 1984.

   J-P. Habicht, J. Mason and H. Tabatabai "Basic Concepts for the Design of Evaluation During Programme Implementation" in D. Sahn, R. Lockwood, and N. Scrimshaw op cit.

3. A. Kielmann, et al., Child and Maternal Health Services in Rural India: The Narangwal Experiment, vol. 1 Integrated Nutrition and Health Care, John Hopkins Univ. Press: Baltimore, 1983.

4. R. Bairagi, B. Edmonston, and A. Hye, "The Influence of Nutritional Status on Age Misstatement for Young Children in Rural Bangladesh," submitted to Human Biology, 1984.

5. G. Snedecor and W. Cochran, Statistical Methods, sixth ed., Iowa State Univ. Press: Ames, IW, 1967, pp. 531-534.

6. P. Levy and S. Lemeshow, Sampling for Health Professionals, Wadsworth: Belmont, CA, 1980, ch. 12.

REFERENCES FOR CASE STUDIES REVIEWED IN BACKGROUND NOTES FOR WORKSHOP

Swaminathan, M.C., Rao, N.P., Sastry, J.G., Mathur, Y.N., Sarma, K.V.R., Rao, D.H. and Ramnath, T. (1983) Special Nutrition Programme in India An Evaluation Study. National Institute of Nutrition, Hyderbad, India.

Gilmore, J.W., Adelman, C.C., Meyer, A.J. and Thorne, M.C. (1980) Morocco: Food Aid and Nutrition Education Agency for International Development, Washington, D.C. (A.I.D. Project Impact Evaluation Report No. 8).

Blumfeld, S., Flores, M.B., Mack, M., Norton, M. and Pooley, R. (1982) P.L. 480 Title II. A Study of the Impact of a Food Assistance Program in the Philippines. Agency for International Development, Washington, D.C. (A.I.D. Program Evaluation Report No. 6).

Echenberg, D.F. and Stubbs, H.A. (1984) Working Document: Evaluation of PL 480 Title II and PPNS Program, Senegal. International Science and Technology Institute, Inc., Washington, D.C. (AID/FVA/PPE and USAID/Senegal Contract No. PDC-0800-C-00-3027-00).

Drake, W.D., Gunning, J.N., Horwitz, A., Miller, R.I., Rice, H.L. and Thenabadu, G. (1982) Nutrition Programs in Sri Lanka Using U.S. Food Aid (An Evaluation of the P.L. 480 Title II Programs). Community Systems Foundation, Ann Arbor, MI.

OTHER REFERENCES FOR EVALUATION METHODS

M. Anderson, J. Austin, J. Wray, and M. Zeitlin, Nutrition Intervention in Developing Countries: supp. vol. 1, Supplementary Feeding, Oelgeschlager, Gunn and Hain: Cambridge, MA, 1981.

J. Austin, et al., Nutrition Intervention in Developing Countries: supp. vol. 5, Integrated Nutrition and Primary Health Care Programs, Oelgeschlager, Gunn and Hain: Cambridge, MA, 1981.

G. Beaton and H. Ghassemi, "Supplementary Feeding Programs for Young Children in Developing Countries", Amer. Jour. of Clinical Nutrition, v. 35, 1982, pp. 864-916.

M. Borus, C. Buntz, and W. Tash, Evaluating the Impact of Health Programs: A Primer, MIT Press: Cambridge, MA., 1982.

D. Casley and D. Lury, Monitoring and Evaluation of Agriculture and Rural Development Projects, John Hopkins Univ. Press: Baltimore, 1982.

L. Chen and R. Feilden, "Measuring Impact Using Clinical, Morbidity, and Mortality Data", in D. Sahn, R. Lockwood, and N. Scrimshaw, op. cit.

D. Chernichovsky, "The Economic Theory of the Household and Impact Measurement of Nutrition and Related Health Programs", in R. Klein, et al., eds. Evaluating the Impact of Nutrition and Health Programs, Plenum Press: New York, 1979.

W. Drake, R. Miller, and M. Humphrey, Analysis of Community-Level Nutrition Programs, vol. I: Final Report, Community Systems Foundation: Ann Arbor, MI, 1980.

J-P. Habicht and W. Butz, "Measurement of Health and Nutrition Effects of Large-Scale Intervention Projects", in R. Klein, et al., op. cit.

A. Kielmann, C. Ajello, and N. Kielmann, "Nutrition Intervention: An Evaluation of Six Studies", Studies in Family Planning, v. 13, 1982, pp. 246-257.

R. Martorell, "Measuring the Impact of Nutrition Interventions on Physical Growth", in D. Sahn, R. Lockwood, and N. Scrimshaw, op. cit.

J. Mason, J-P. Habicht, H. Tabatabai, and V. Valverde, Nutritional Surveillance, World Health Organization: Geneva, 1984. Chapter V, "Nutritional Surveillance for Programme Management and Evaluation."

APPENDIX I

LIST OF PARTICIPANTS

Dr. Michael Dibley
Division of Nutrition
Center for Health Promotion and Evaluation
Centers for Disease Control
1600 Clifton Road, Northeast
Atlanta, GA 30333

Dr. William Drake
Community Systems Foundation
1130 Hill Street
Ann Arbor, Michigan 48104

Dr. Barry Edmonston
Associate Professor of Sociology
Cornell University
Ithaca, NY 14853

Ms. Judith Gilmore
USAID Bureau for Food for Peace
and Voluntary Assistance
Office of Program, Policy and Evaluation
Department of State
Washington, D.C. 20523

Dr. John Haaga
Cornell Nutritional Surveillance Program
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Dr. Jean-Pierre Habicht
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Ms. Chessa Kenrick
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Ms. Kathleen Kurz
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Dr. Reynaldo Martorell
Food Research Institute
Stanford University
Stanford, CA 94305

Dr. John B. Mason
Cornell Nutritional Surveillance Program
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Dr. John McKigney
Office of Nutrition
Bureau of Science and Technology
Agency for International Development
Department of State
Washington, DC 20523

Dr. Philip Nieburg
Division of Nutrition
Center for Health Promotion and Evaluation
Centers for Disease Control
1600 Clifton Road, Northeast
Atlanta, GA 30333

Dr. Francesca Proja-Ronchi
Food Policy & Nutrition Division
Food and Agricultural Organization
Via delle Terme di Caracalla
00100 Rome, Italy

Ms. Rhonda Sarnoff
Regional Nutritionist
Catholic Relief Services
P.O. Box 48932
Nairobi, Kenya

Dr. David Saunders
Lecturer in Child Health
University Hospital
Harare, Zimbabwe

Ms. Hope Sukin
USAID Bureau for Food for Peace
and Voluntary Assistance
Office of Program, Policy and Evaluation
Department of State
Washington, D.C. 20523

Ms. Karen Test
Cornell Nutritional Surveillance Program
Division of Nutritional Sciences
Cornell University
Ithaca, NY 14853

Dr. Victor Valverde
Food and Nutrition Policy Division
Instituto de Nutrition de
Centro America y Panama (INCAP)
AP Postal 1188
Guatemala City, Guatemala

Dr. Ken Williams
Regional Advisor in Statistics
UNICEF
East Africa Regional Office
P.O. Box 44145
Nairobi, KENYA

WORKSHOP ON METHODS FOR EVALUATING IMPACT OF CHILD FEEDING PROGRAMS

Cornell University

(6-7 September 1984)

THURSDAY, SEPTEMBER 6

9:00	Welcome and Introduction	Martha Van Rensselaer Hall Room 114
	Discussion of Agenda Formation of Working Groups	
10:00 - 10:15	Coffee break	
11:00	Working Group Discussions (proposed, to be finalized in first session)	

A. **Design** — MVR, Room 142

- evaluation objectives
- clarifying program objectives
- study designs (prospective vs retrospective, cross-sectional vs longitudinal, comparison groups)
- comparisons (with/without, time in program, variation in delivery/components, natural experiments)

B. **Outcome Indicators** — MVR, Room 114
(need, feasibility, methods for analysis)

- anthropometry, (to include growth velocity; use of growth charts; problems with aging, etc.)
- morbidity
- mortality
- food consumption
- other outcome
- controlling variables

C. **Confounding** — Savage Hall Annex
(implications for study design, and analysis)

- dropouts
- selection biases
- controlling for non-comparability in analysis
- aging, regression to means, secular trends, seasonality

12:00	Lunch	
1:00	Working Group Discussions	
	A	MVR, Room 142
	B	MVR, Room 114
	C	Savage Hall Annex
5:00	Adjournment	
6:00	Reception (transport from trailer or motor lodge at 5:45 PM to 112 Warwick Place)	Mason home

FRIDAY, SEPTEMBER 7

8:30	Presentation and review of working group conclusions	MVR, Room 142
10:00	Coffee break	MVR, Room 142
10:30	Working Group Discussions	
	D. Presentation of impact evaluation results: for whom, how, level of detail, how to review findings and conclusions	MVR, Room 142
	E. Adequacy evaluation - when, how: assessing targetting, delivery, gross outcome	Savage Hall Room 136
	F. Procedure for planning future evaluations	Savage Hall Annex
12:00	Lunch - meet students	Savage 130
1:00	Complete working group discussions	MVR, Room 142
	Review working group conclusions and discuss recommendations	
5:00	Close	
	(Tea at 3:00 PM)	