PD-AA-9468
ISN = 24885

6-2

CORNELL NUTRITIONAL SURVEILLANCE PROGRAM

Working Paper Series

Division of Nutritional Sciences
New York State Colleges of Human Ecology & Agriculture and Life Sciences
Statutory Colleges of the State University of New York
Cornell University
Ithaca, New York 14853

Cornell University is an equal opportunity, affirmative action employer and educator.

PN AAQ 188

NOTE ON A FRAMEWORK FOR MONITORING AND EVALUATION OF UNICEF/WHO NUTRITION IMPROVEMENT PROGRAMMES

by

J.B. Mason and J.G. Haaga

March 1983 **No.16**

The CNSP is supported by Cooperative Agreement #AID DSCAN 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 Unversity Agricultural Experiment Station.

-1-

This is a draft note on suggestions for a possible framework for monitoring and evaluation of the proposed programmes to be supported by UNICEF/WHO. It responds to a request made at a meeting in early February 1983 at UNICEF headquarters. The note is in four sections. Section I outlines our understanding of the general considerations involved for individual programmes. Section II gives a hypothetical example of some of the suggestions being put forward, based on specifications available for one proposed program. Section III begins to outline some practical steps that could be applied in the design of suitable monitoring and evaluation systems. Section IV gives further details on considerations for designing evaluation of net impact -that is, ascertaining the relationship between programmes and the outcome obtained. The document is intended for discussion, and could be turned into a guideline to assist the design of such systems.

The Joint UNICEF/WHO Support Program for the Improvement of Nutrition (1982-1986) is described in broad terms in the Executive Board paper of March 1982 (ECOSOC, 1982). Several points are important for the present purposes, and are briefly summarized here, taken from the ECOSOC paper - (numbers in brackets refer to paragraphs in that paper).

The impact objectives(12) of the overall five-year program are stated as:

- reduction of infant and young child mortality and morbidity;
- better child growth and nutrition;
- improvement of maternal nutrition.

The operational objectives(13), in brief, are:

- to build national capacity for improving nutrition, through (a) nutrition components in PHC, and (b) including nutrition objectives in other sectors' activities, with health sector help;
- advocacy
- international cooperation;
- information.

Some of the programs are defined as:

- in PHC(28): activities aimed at breastfeeding, weaning practices, growth monitoring, rehabilitation, iron-folate and Vitamin A distribution, maternal nutrition.
- in other sectors(41): food interventions aimed at: national policies, food aid, food production subsidies, consumption subsidies
- income-related activities(44) - i.e. nutrition in development policies and programs.

WHO/UNICEF will be directly involved in at least 5 country projects(43). At least 10 case studies are aimed at(50), importantly not as pilot projects, but "on operational activity on a sufficient scale to illustrate the problems of national services", e.g. at least one administrative unit of the country, with a population of the order of 500,000 - 2,000,000.

The interpretation of this plan, to us, is that the monitoring and evaluation component of the overall plan is needed for several related purposes. First, it is essential to use this opportunity for establishing/supporting operational-scale programs which bring about significant improvements in nutrition, and an information system is required for proper management of these programs. But the objectives go beyond this, because these programs need to be seen to work, and their impact estimated. Impact means net effect - due to the program - and with estimates of cost can allow estimation of impact per unit cost, which is cost-effectiveness. Further, information is needed to introduce nutritional considerations into non-nutritional activities in other sectors. Finally, the information needs to be available for dissemination, as case-studies etc.

In our terminology, that is in the context of nutritional surveillance (ACC-SCN 1981; National Academy of Sciences 1982), this means that we must consider methods of nutritional surveillance for program management, primarily for "adequacy evaluation" built into programs in order to help these programs succeed (Mason et al. 1983 Chapter 4). But some estimates of net effects, or impacts on nutritional outcome, will also be needed, which requires more careful design, data collection, and analysis on a subset of projects and/or project participants. Finally, some information is also needed for setting nutritional objectives outside direct program delivery, hence for analysis of non-nutritional policies and programs. This latter requirement is not explicitly addressed in this note, but is referred to in other publications (e.g. Mason, 1982). The methods and thinking referred to below have been described, in the context of nutritional surveillance in other CNSP documents which are referenced in the text.

The suggestions here apply to designing a monitoring and evaluation system for individual country programs. Evidently the overall system will be built up from these. However, the overall system should also be kept in mind, as eventually the sum of knowledge aimed for should result from a synthesis of

national and local results. First, we need to know what works under what circumstances. For this, a good spread of focussed evaluations of impact on small parts of projects would seem to be desirable because impact evaluation may not be feasible in all countries.

I. GENERAL CONSIDERATIONS FOR INDIVIDUAL PROGRAMMES

A. **Objectives of Program.** These are taken to be to:

1a. Effectively deliver certain appropriate goods and services to certain people.

1b. Learn from this to do programmes more effectively and more widely.

2a. Improve health and nutrition for beneficiaries

2b. Know that health and nutrition are improving.

3. Use the experience to design programmes that are more cost-effective in terms of health and nutrition outcome, for future wider application.

1b, 2b, and 3 mean that monitoring and evaluation must be included in the program.

B. **Objectives of Monitoring and Evaluation.** These are taken to be to:

1. Ensure good management of individual programmes.
2. Learn lessons with respect to management.
3. Ensure that health and nutrition improve at an adequate rate for beneficiaries of individual programmes.
4. Evaluate impact of the programmes.
5. Evaluate impact of individual components of the programmes.

Not all these objectives of monitoring and evaluation are likely to be appropriate for all programmes. It is likely that 1 and 3 are important for most programmes. 2 would be important for part of the programmes. 4 and 5 are important for selected programmes, or parts thereof (in particular, those for which there is not a large base of field experience on which to draw, or whose effects are expected to vary widely across populations).

C. Planning

Design of monitoring and evaluation depends upon attention to design of the program itself. Design of a program likely to be effective involves:

- having an adequate base for thinking:
  - (a) that a particular intervention can be effective (i.e. reach objectives) in principle
  - (b) that the intervention can be effective in the existing circumstances - i.e. it is appropriate and feasible
  - (c) that a particular intervention is a better choice than an alternative intervention
- correctly identifying those who need the program, can be brought into the program, and are likely to respond to the components of the program
- knowing the level of resources for specified beneficiaries required to meet the program's objectives, and planning to ensure this delivery.

D. Planning requirements.

The suggestions here are necessary both for planning the program, as well as for successfully designing monitoring and evaluation.

1. Within each program (with the objective of ensuring a successful program in terms of delivering appropriate goods and services at the right levels of input to the designated people (the people in need, likely to respond), it is necessary:
   - to identify appropriate outcome objectives, with time frame
   - to identify causal factors of the problem to be solved, that are open to feasible and effective intervention.
   - to identify options for delivering goods and services to tackle these causal factors.
   - to choose between these options on basis of likely cost effectiveness (i.e. impact per unit cost), feasibility and simplicity in management.
   - to design an information system to ensure effective management, with periodic checks on the adequacy of outcome.
2. Within a part of each program, two different more intensive efforts are suggested. The first (a) uses much the same information as given

in 1 above, and differs mainly in management. The second (b) has a quite different intent, to estimate whether changes in outcome (e.g. improvement in nutrition) are causally related to the program itself. These two more intensive efforts are:

(a) To allow flexible and adaptive management. This requires a decentralized budget and capable local leadership - the intention is to allow more rapid evolution of the program and problem solving.

(b) To estimate impact. This requires some form of comparison, collection of additional data, and more depth in analysis.

The rest of this note refers to these three levels of evaluation/management sequentially, as types 1, 2, and 3. Type 1 is referred to as "Routine Adequacy Evaluation", type 2 is referred to as "Flexible Management/Adequacy Evaluation", and type 3 is referred to as "Impact Evaluation".

E. **Theory for this monitoring and evaluation.**

Three levels of effort are required, to summarize:

1. Process and minimum outcome adequacy evaluation, for helping effective routine management of all programmes (Routine Adequacy Evaluation).
2. Process and outcome adequacy evaluation, with strengthened management, and decentralized decision making (Flexible Management/Adequacy Evaluation).
3. Carefully designed (usually quasi-experimental - see ref. 1, Table 1.6) evaluation to estimate impact; this may also include intermediate outcome objectives (e.g. changes in KAP); it requires comparison groups and/or variation in the level of delivery (Impact Evaluation).

1. Routine Adequacy Evaluation:

- monitor overall delivery: Is delivery in line with the plan (assuming the plan itself was sensible)? Delivery is measured in physical units or number of service contacts, and, if possible, expenditure per head.

- monitor targetting -(see 2 x 2 tables in ref. 1 Table 1.8). This gives indicators of coverage, focussing, leakage, and delivery.
- periodically monitor changes in outcome indicators (measures of health and nutritional status) relative to objectives. If feasible a minimum adequacy chart (control chart) needs to be set up, as in ref. 2 figure 4.2.
- modify delivery of planned goods and services in response to this information.

2. Flexible Management/Adequacy Evaluation:
   - The basic theory is the same as no. 1 above, but requires that management is able to evaluate the information and make on-the-spot decisions to adapt the program in the light of this information.
   - It is assumed that in routine programmes the management may not always be authorized or able to make such decisions.
   - Other information on process and outcome, obtained not only from program contacts but also, for example, by survey, may be added in.

3. Impact Evaluation:
   - The objective here is to allow some conclusions on whether there is a causal link between the program (or program components) and improvement in outcome.

   Method (see section IV for more details): some form of comparison is needed, either or both of with/without program, before/after program. The comparison can be done using varying levels of program delivery (provided this is not completely confounded by other factors, such as accessibility). The comparison groups do not need to be precise controls, i.e. randomized or equivalent in every respect except for the operation of the program, provided other factors associated with outcomes are measured. This sort of analysis requires research institute capabilities (which may often be found in-country - however, technical assistance at design and analysis stages will often be needed).

F. Practice of Monitoring and Evaluation

1. Routine Adequacy Evaluation.

The reporting system mainly covers process, with outcome data obtained primarily through program contacts.

Process covers:

(a) whether there is delivery of correct (or adequate) quantities of goods and services per time to project area

(b) whether goods and services then reach specified target groups (see ref. 1 Table 1.8; ref. 2 Table 4.2 and 4.3 and accompanying text).

Outcome covers:

Whether periodic estimates of outcome indicators (e.g. nutritional status) for (a) recipients and/or (b) targetted population show an adequate trend. This depends on suitable specification of objectives, e.g. declining prevalence of malnutrition.

The system would draw on:

- Overall delivery from administrative reports of delivery of supplies to area; staff employed; etc.
- Delivery to target groups: from program contacts, requires recording and tallying system in clinics, etc., which includes definition of whether recipient is targetted. This is only relevant for some services - for example, health care may be for everybody in the area, but supplementary foods only for those children with a certain degree of malnutrition.
- Outcome indicators (e.g. morbidity, nutritional status) would be obtained from program contacts only: e.g. by recording and tallying proportions of children above/below weight/age cutoff (e.g. 80%). This would probably be reported monthly, but assessed every three months.

Hand tallying and simple tabulation are all that are likely to be needed for analyzing these data.

Management decisions from these results.

1. Process, concerning overall delivery: identifying problems, solving these locally or reporting on up the administrative structure.
2. Process, targetting: tighten up selection and outreach procedures; assess whether the plan is feasible and if not, refer up.
3. Outcome: Inadequate outcome leads first to re-examining 1 and 2. If these are okay then either planned objectives were unrealistic or unexpected adverse conditions have affected outcome. Either way, the program needs review at a higher level.

2. Flexible Management/Adequacy Evaluation

The data side is much the same as for Routine Adequacy Evaluation. The management, however, would have increased resources to respond autonomously to the information -itself replanning, e.g. by changing specification of activities, target groups, schedules, etc. This decentralization could allow much more rapid evolution of an effective (information - based) design and management.

3. Impact Evaluation

This is carried out only on a sub-set of project activities (e.g. by area). It may be decided not to do this in all countries concerned. The objective is quite different from 1 and 2 - to investigate whether there is likely to be any causal relation between the program and changes in outcome (e.g. nutritional status). This will allow then some estimates of the cost-effectiveness of the program, providing adequate records are kept of the costs involved in the part of the program that is being so evaluated and providing shared costs are estimated and realistically allocated among program components. The main differences from evaluation of types 1 and 2 are the inclusion of comparison groups, the need for collection of additional variables, and the greater analytical depth.

Sources of data, possibilities are:

- from service delivery points (e.g. clinics) before program, and/or without program, and/or with lower delivery of program, as well as for full participants. This data is to be obtained through program contacts.

- from interviews/measurements in households/villages specially selected in a sample survey. This requires either that program staff have a capability for survey work, or that enumerators are hired specially for the exercise, or that the program area falls within the sample frame of a suitable wider survey.

Approximately annual (taking account of seasonality) data collection and analysis are required. In contrast to types 1 and 2, computer - based analysis is likely to be needed.

II. EXAMPLE OF SOME ASPECTS OF A MONITORING AND EVALUATING SYSTEM

The purpose of this section is to give some suggestions for specification of target groups, delivery rates, process and outcome indicators, for certain components that might reasonably be contained within one of the programmes under consideration. Details are given in table 1. The actual figures are taken from one example project document. The specifications, generally derived from that project document, are given in Table 2. The table is arranged by component.

1. **Primary Health Care (say equivalent to dispensaries)**

   The target group could be 25% of the population in the area, defined for example as the worst served, or those most in need; need could be defined in terms of health, or nutritional status, or some other equivalent measure. The delivery rate (calculated from Table 2) is of 60 primary health care workers per year, with a given number of expected contacts per year. Indicators of delivery and targetting are simply in terms of the number of primary health care workers stationed, and whether they are stationed in the targetted areas. The outcome indicators are more difficult. It is probably not possible to measure changes in morbidity through contacts at a dispensary, although it may be possible to measure the nutritional status of those coming to the dispensary. It would be expected that the underlying health or nutritional status should improve.

2. **Expanded Program of Immunization**

   The targetting, delivery, and process indicators are straightforward, as

shown in the table. Outcome indicators would depend on reporting of incidence of diseases; again, this can be very unreliable as a measure of true incidence in a population. For some diseases, specilized outcome indicators are feasible (e.g. lameness surveys for poliomyelitis).

3. **Nutrition component**

   Three subcomponents are considered: growth charts, distribution of vitamin A capsules, and nutrition education. Target groups may be defined as shown in the table, and delivery rates and process indicators are fairly straightforward in principle. Outcome indicators are similar, in this example, to those expected from primary health care. As for the other components, it is useful to categorize these into adequacy and impact. Adequacy in this context refers to whether the health and nutrition conditions for recipients are adequate; impact attempts to ascribe this outcome to the program.

4. **Household food production and storage**

   Considerations here are similar to the other components as shown in the table. One difference, however, is that the impact (as opposed to adequacy) of this component would require considerably more detailed study (an example is discussed below in Section IV).

5. **Supplementary feeding**

   This component requires screening. This imposes more stringent conditions on the monitoring, but on the other hand would be expected to have more easily detectable impact if it really was an effective component.

TABLE 1

EXAMPLE OF MONITORING SYSTEM FOR FIVE POSSIBLE COMPONENTS OF A NUTRITION IMPROVEMENT PROGRAM *

Component	Planned Target Group	Planned Delivery Rate	Process Indicators: Delivery	Process Indicators: Targetting	Outcome Indicators (Adequacy/Impact)
1. Primary health care (say = dispensaries)	The 25% (a) worst served (b) in most need	60 PHC workers per year (contacts: 30,000 first year, etc. . . .)	(a) No. PHC workers stationed per year (b) No. of contacts	(a) Stationing as planned in priority areas (b) Contacts from target	Morbidity: not measurable as prevalences or incidences, nor probably as severity. Conceivably as outcome e.g. mortality. Can measure how 'healthy' patients are by e.g. nutritional status. Village birth and death registrars may be feasible under some circumstances.
2. EPI	Total population of children and women in child-bearing years	100,000 people (or 20,000 children) per year	-Delivery of vaccine -Numbers vaccinated -IHI survey examining vaccination records or scars	- - - -	Reporting of incidences of diseases vaccinated against, specialized surveys (lameness, measles reports)
3. Nutrition					
- Growth charts	As PHC, all children in this group	6,000 per year (used)	Distribution of growth charts	(a) Distribution by area (b) Presentation of filled in growth charts	Indistinguishable from 1 & 2: Adequacy: growth adequate Impact: e.g. cf. wt/age of those with growth charts vs new entrants (possibly controlling for age, distance, etc.)
- Vitamin A capsules	(a) All children with growth charts	6,000 per year	Distribution of Vit. A capsules.	Distribution by target area	Signs of Xerophthalmia
or	(b) All children in affected areas	?			
or	(c) All children, as EPI	20,000 children per year			
- Nutrition Education	Say used for those less than 80% W/A	At 30% prevalence, 1,800 per year, =	Dist. of materials (spot check of uses?)	(a) By area (b) Reporting on W/A of e.g. those home-visited	(a) Adequacy: growth chart weight/age of those contacted; ? KAP survey. (b) Impact: e.g. cf. with those with less frequent home visits
HH food prod. & storage	In PHC areas, all hh's	Say 150 villages, = 30 villages per year	Hh's contacted, supplies (e.g. seeds) delivered	Services, contacts by area	(a) Adequacy: indistinguishable from PHC (b) Impact: needs a special study
Supplementary feeding (children at risk)	Children 80% W/A in PHC areas	@ 30% prevalence = 1,800 at any time, say x 2 = 3,600 p.a.	Delivery and utilization of supplies	Recording W/A of those to whom supplementary food delivered, recording W/A of others, to detect those not receiving supp. foods	(a) Adequacy: W/A, and growth of those with supp. foods (b) Impact: (i) cf. those (for some reason) not receiving food, of 80% W/A (ii) cf. those say 80 to 85% W/A, etc.

TABLE 2

BASIC DATA ASSUMPTIONS FOR TABLE 1

(These are taken directly from an example project document)

Area: 0.6×10^6 people

Expenditure: $\$2 \times 10^6$ p.a. of which approx 50% is goods and services, i.e. approx. $\$1 \times 10^6$ per, say, 0.5×10^6 people p.a. = $2 per head per year

Probably need about $5 per head of target population per year for "worthwhile" effect.

Therefore need to target at maximum 40% of population

Stated Objectives:

- Establish large scale working example of services to reduce: hunger
  malnutrition in young children and mothers
  child mortality
  inadequate child development

Target population: 150,000 = 25% of total population. Gives approx. $7 of total target population per year.

Components:

1. Primary health care, 1 PHC worker per thousand population
2. EPI
3. Nutrition growth charts; Vitamin A capsules, nutrition education materials
4. Household food production and storage
5. Supplementary feeding of malnourished children

N.B. Costs per component cover different recipients. For example, immunization for pre-school children and women of child-bearing age might cost $7 per head. The target population for this component is a sub-set of the total target population, e.g. 30% of 150,000 = 45,000; at $7 per head = $315,000 in the first year, less thereafter.

III. STEPS IN DESIGNING MONITORING AND EVALUATION SYSTEM

(Sections match Part I)

A & B. Review Overall Purposes of Program and its Monitoring and Evaluation.

Determine priorities, balance between e.g. types 1 to 3 evaluation, approximate resources available for each program objective and monitoring and evaluation exercise. (See also Stage 1 in Ref. 1, Table II.1)

C & D Planning the Program and Monitoring and Evaluation.

- See Stage 2, Ref. 1, Table 2.2.
- Is there an adequate basis for the program design?
- Can target groups be identified operationally?
- Are delivery levels determined and likely to achieve objectives?
- See other questions in Part I, D.

E. Theory, Translating to Practice.

- Define possible management decisions, and administrative levels at which they can be taken, responding to information from the monitoring and evaluation system.
- Set up dummy outputs, periodicity, proposed reporting mechanisms.
- Design reporting forms.
- Design tallying forms.
- Determine personnel availability.
- Field test forms.
- Propose training for data collectors, transmitters, and analysts.
- Set up criteria for adequacy of delivery and targetting (focus, coverage, etc.)
- Set up control chart for adequacy of outcome. Ideally this requires calculation of proposed reduction in prevalence of e.g. malnutrition, and hence implicit gross cost-effectiveness. In practice, a best guess of e.g. malnutrition reduction from say 30% prevalence at beginning to 20% prevalence at end in target group, already gives the basis for a useful start. This can then be translated into a rate of change chart.
- Propose procedures for management decisions based on data outputs, i.e. from information on delivery, targetting, and outcome.

This procedure should be iterative, and a re-examination of the preceding steps in this section is now needed.

- Setting up impact evaluation requires identification of possible comparison groups, confounding variables, and hence, a more extensive data collection system, as well as analytical capability.

IV. IMPACT EVALUATION

1. **Purpose** - to estimate net effect of program on outcome indicators. Presupposes, but goes beyond, Routine Adequacy Evaluation, to investigate the links in a causal chain connecting program activities with changes in health and nutritional status of intended beneficiaries. That is, goes beyond asking "Did activities happen as planned?" and "Are outcome indicators satisfactory in beneficiary population?" Information on effectiveness is potentially useful for decisions about expansion or replication of a pilot program. If combined with cost information (cost-effectiveness analysis), it can be used to set priorities, and to support resource allocation decisions at fairly high levels: among components of PHC, for example, or between PHC and rest of health sector. The audience for impact evaluation overlaps but it is not exactly the same as for adequacy evaluation. (1 and 2) In the UNICEF/WHO case, assessment of impact of different elements of the overall programme is crucial.

2. **General Method** - Evaluations try to compare the situation with the program or program component to what the situation would have been in the same population without the program. The latter is impossible to measure directly. Various research designs can give better or worse approximations, leading to greater or less confidence in results. Evaluators can never prove X caused Y, but they can strengthen their inference about the program's effect by taking into account competing explanations for observed changes. In an experiment (strictly defined), subjects are assigned randomly to treatment or control groups. This is not usually feasible for field interventions (as opposed to clinical trials). Quasi-experimental designs are more likely to be used in evaluating the components of pilot PHC programs.

3. **Quasi-Experimental Research Designs**

Note: In the following abbreviated descriptions of research designs,

0 = measurement of outcome indicators, with or without other variables

X = program actually begins operation. For further discussion, see Mason, Habicht and Tabatabai (1982), table 1.6 and pp. 71 to 73.

A. Comparison with baseline survey on the same population.

Treatment group: $0_1 \rightarrow X \rightarrow 0_2$

This allows virtually no inferences on impact. The major problem is that too many other things change, while the program is going on. It is difficult to ascribe causation. Also, this research design won't pick up program effect "masked" by other trends. The health status of the population may be the same or worse at point 0_2, despite a positive net impact of the program, if the program served to prevent worse deterioration. Evaluations must take into account migration, and selection biases which would make the "before" and "after" group not truly comparable. Seasonal variations or natural disease cycles are often mistaken for program effects. Any effort to assess net impact should attempt to go beyond this design.

Repeated measurements can be made before and/or after starting the program (trend analysis).measurements are made before and/or after starting the program. $0_1 \rightarrow 0_2 \rightarrow 0_3 \rightarrow X \rightarrow 0_4 \rightarrow 0_5 \rightarrow 0_6$

If indicators show a clear break in health or nutritional status trends after initiation of the program, the inference that the effect is due to the program is stronger than comparisons with the baseline alone would allow. One still can't uncover "masking" effect of adverse changes in non-program factors, and migration and selection biases may still be operating, though their detection may be facilitated by this kind of monitoring.(Referred to as 'interrupted time-series in Mason etal. 1982, p. 71).

B. Comparison with national or regional averages.

Treatment group: $X \rightarrow 0_2$

Rest of population: $0_1 \rightarrow 0_2$

This design is frequently used, especially in informal evaluations. Before making an inference about program effects, though, one would need to know

or to estimate how comparable the two groups are that are being compared. It is easy to mistake a program effect for a difference in health status caused by non-program factors.

C. <u>Use of comparison group receiving no (or different levels of treatment).</u>

Treatment group: $0_1 \rightarrow X \rightarrow 0_2$

Comparison group:

$0_1 \rightarrow 0_2$

This design is often administratively difficult, but it can lead to strong inferences, depending on how similar groups were to start with, and how well changes in other factors related to health/nutritional outcomes can be monitored. Even if the groups are not exactly the same in levels of income, landholding, access to clean water, other services, age distribution, etc., or if they are differentially affected by environmental changes not related to the program, these differences can be partially controlled for in the analysis, if they are measured.

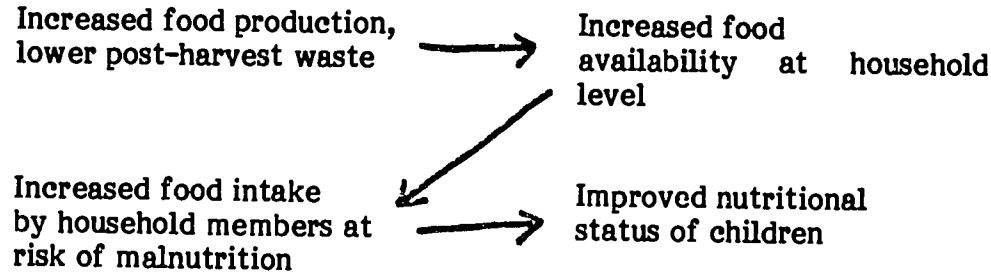
Comparison with incoming groups is a variant of C in which the comparison groups are new areas to which the program is being expanded. Most programs do not start everywhere at same time so baseline or first-contact data from the new areas can serve for comparisons with areas in which the program has been established longer. Again, measurement of other, non-program factors affecting health outcomes will strengthen the comparison.

Alternatively, in many situations program delivery itself may vary naturally. If this can be measured (e.g. number of patients seen, supplies delivered) the association between delivery and outcome can be investigated. Moreover, in programs not start gradually, the first measurements (outcome and delivery) taken with the program start may be able to substitute for the baseline survey.

4. **Example**

Evaluation of Impact of Household Food Production and Storage Program on Nutritional Status (Component No. 4 in Table 1, Section II above)

A. Hypothesized causal links:



B. Research design: Compare villages where the activity is taking place to incoming villages in succeeding years.

The unit of analysis will be the village, on the assumption that this particular program component is targetted to entire villages rather than households.

Timing of Measurements:

Group	No. of villages	Year 1	2	3	4	5
A	30	A1	A2	A3	A4	A5
B	30		B2	B3	B4	B5
C	30			C3	C4	C5
D	30				D4	D5
E	30					E5

(Underlined entries indicate baseline surveys)

The effect of one year of program activities can be estimated by comparing:

A2 vs A1	also	A2 vs B2
B3 vs B2		B3 vs C3
C4 vs C3		C4 vs D4
D5 vs D4		D5 vs E5

Such a design allows larger sample sizes than before - after comparisons alone. It also makes it possible to control for extraneous events like an unusually poor harvest in one year, an epidemic, etc. If no baseline surveys are available, one can still estimate effects by comparing B3 vs A3, C4 vs B4, D5 vs C5, etc. Also one can look for longer-term impacts: A4 vs C4, A5 vs C5 and

D5, etc. The variation across 150 villages in the intensity and duration of program activities allows estimation of a "dose-response" curve: Do greater inputs always lead to improved outcomes? Under what circumstances? By how much? Up to what point of diminishing returns?

Potential confounding (competing explanations for differences observed in simple comparisons).

1. Accessibility - If the 30 villages in group 1 got services first because they were the most accessible (and accessibility is related to other factors that affect nutritional status- income from crops, availability of medical care), the single comparisons will over-state the program impact, if any.
2. Program or recipient selection bias - If the 30 villages in group 1 are somehow selected as those where service most needed, or if the most needy villages in each group are more likely to participate, then simple comparisons will under-state the true program impact. However, more commonly self-selection in the direction of over-estimating impact, because the self-selected participants are those most likely to improve.
3. More generally, non-comparability (in level and sources of income, age distribution and pre-program health status of population, etc.) of villages in the different groups will bias the estimation of program impact. This bias could be in either direction, exaggerating or masking the net impact of the program.
4. Differences in participation, and levels of program activities among and within groups of villages complicate evaluation studies. A very common problem is simply specifying what actually happened in the "treatment" villages, that is, what exactly is being evaluated. Especially when management decisions are decentralized, the evaluations need to be careful to distinguish between "X was tried but didn't have an effect" and "X was not in fact tried".

Simple between group comparisons are greatly strengthened if confounding factors can be accounted for (partially) in cross-tabulations or multivariate analysis.

In the case of a program which is reaching primarily the worst-off villages or households, then controlling for non-program variables may reverse the algebraic sign of a simple 2-way comparison of program and non-program groups. This applies with particular force when screening for selection into a program is used. The program participants should appear worse-off on simple comparison with non-participants. As another example: Suppose villages after two years in household food production and storage program still have a higher overall prevalence of underweight children than a comparison group of villages just entering the program.

Simple Comparison of Outcome Indicators:
% children 80% weight-for-age

Group 1 (2 years on program)	29% (30 villages, 1300 children)
Group 3 (just entering)	27% (30 villages, 1300 children)

But the villages in Group 1 have less access to adequate, clean water (and this is a presumed greater incidence of diarrheal diseases)

Comparison of Outcome Indicators Controlling for Water Supply:
% children 80% weight-for-age

	Water from standpipes (clean)	Water from canals (dirty)	TOTAL
Group 1	20% (n = 130)	30% (n = 1170)	27% (n = 1300)
Group 3	25% (n = 1040)	35% (n = 260)	26% (n = 1300)

The inference that the program brought about a 5% difference between Group 1 and Group 3 would be strengthened if the intermediate step (incidence of diarrheal disease) could be compared; and especially if similar results appear when other factors that affect attained weight-for-age are controlled.

Regression toward the mean complicates analysis if program participants selected by low values on a screening test. Statistical controls can only partly take the place of basic similarity between the treatment and comparison groups.

In the case of a program that doesn't reach the worst-off (because the truly destitute are unable to participate, or because it is administratively difficult to reach the destitute until after the program is well under way), controlling for non-program factors affecting growth similarly strengthens any inference about program effect (positive or negative) or lack of it. This time, the difference in outcome indicator between groups is less pronounced as more confounding factors are controlled, but the inference about the program's effect is strengthened as it "survives" multiple controls.

Simple Comparison of Outcome Indicators:
percent of children 80% weight-for-age

Group 1	22% (n = 1300)
Group 2	33% (n = 1300)

Comparison of Outcome Indicators Controlling for Water Supply
% children 80% WA

	Clean water	Dirty water	TOTAL
Group 1	20% (n = 1000)	30% (n = 300)	22% (n - 1300)
Group 3:	25% (n = 300)	35% (n = 1000)	33% (n - 1300)

These are two extreme cases: one where the children in the program villages appear to be worse off than the children in the villages without the program, but the net impact of the program is in fact positive; and the other where the confounding tends to exaggerate, not to mask, the program's impact. In practice, the situation is not always so stark, since some confounding factors may offset others to some degree, and since variation of the confounding factors between groups may not be so great as in these hypothetical examples. (That is, the villages in the comparison group often happen to be broadly similar

to those in which the intervention was first tried, particularly if there was some conscious effort to make them so.) In our examples, the program had a positive net impact, but the same types of confounding could operate either to mask or to exaggerate a net negative impact or to create an illusion of an impact where in fact there was none. Analyses are futher complicated if there are threshold effects of program components or interactions between components, (e.g. if measles immunization only has a strong effect if coverage is more than 80% over a 2-year cycle of transmission, or if the food production/storage program only has effect when combined with land redistribution or nutrition education.)

The variables usually needed include:

1. Factors to be controlled in analyses (e.g. other health services, water supply, household incomes, etc.)
2. Measures of program inputs (what services actually delivered? when?)
3. Outcome indicators (dependent variables in analyses).

Where possible, especially in small-scale studies of new interventions, various intermediate indicators related to steps in the causal chain connecting inputs to outputs will also be gathered. In our example above, intermediate indicators might include household food supplies or even dietary surveys.

5. Problems with impact evaluation

A. Expense - Impact evaluation cannot be instituted on routine basis, like the first two kinds of evaluation. The data and analytic requirements exceed those of adequacy evaluation. Accordingly once it has been established that an intervention does cause improvement in certain populations under certain conditions, administrators then monitor program effectiveness through types 1 and 2 evaluation until conditions change in such a way that the earlier conclusion may no longer hold. The oft-posed question which type of evaluation is more useful is beside the point - it all depends on the stage of the program, the degree of uncertainty about causation, and the decisions to be made. In deciding the level of resources to be devoted to evaluation as a whole and among the different types of adequacy and impact evaluation, program planners need to perform an informal cost-effectiveness analysis of evaluation research itself; some study designs may simply cost more than the expected value of the information to be derived. (The more usual case, however, is that too few resources are devoted to evaluation, analysis, and reporting.)

B. Uniformity of "treatment" - As discussed above, this is a problem especially when management is flexible, or when the intervention consists of a complex bundle of services. Impact evaluation may be feasible only for discrete components of the overall package.

C. The choice of outcome indicators requires specification of the goals of the program and the purpose of the evaluation. Adequacy evaluations are often mistaken for impact evaluations, then the "outcome" indicators used are not actually measures of health status, but are measures of service delivery. The measures used as outcome indicators need to be sensitive enough in the relevant range to detect the impacts expected of a program in the sample size available. Mortality rates, especially cause-or-age-specific mortality rates, often cannot be used because of random fluctuations in small samples. Morbidity rates often suffer from "reporting bias" - the number of cases of infectious diseases reported in clincs or even household surveys may increase, despite decrease in true incidence, if program causes more cases to be detected and diagnosed. Growth rates are usually a good proxy for health/nutritional status of children, but may not be the outcome of interest for some PHC components.

D. Migration into or out of program areas. If those people adversely affected, or unaffected, by the program emigrate, the outcome indicators in the remaining population may not show the total net impact of the program. For example, some types of agricultural extension program may decrease employment opportunities for casual workers or seasonal migrants, or take over commons formerly used by nomads, etc. If the landless, migrants, or nomads are not around to be measured, evaluation studies could miss this effect.

6. **Conclusions** - Impact evaluation happens informally anyway, as part of health policy "folklore". (e.g. "X never works and costs too much" or "Y is an essential part of any PHC program") But planners need more specific guidance from past experiences: Under what circumstances might X work? or How much Y, and for whom? Impact evaluations, with their design and assumptions spelled out, could make experience of the UNICEF/WHO programs interpretable (and cumulative), - In the long-run one of the most important outcomes would be to finally give solidly based guidance on what works, under what circumstances, and how much does it cost.

References

ACC-SCN (1982) Report of the International Workshop on Nutritional Surveillance held in Cali, Colombia, 14-17 July 1981. SCN 82/10. Rome : ACC -SCN.

ECOSOC (1982). Joint WHO/UNICEF Support for the Improvement of Nutrition. Proposed Five-year Programme of Work 1982-1986. E/ICEF/L. 1441. 31 March 1982. Economic & Social Council. United Nations.

Mason J. B. (1982) Minimum Data Needs for Assessing the Nutritional Effects of Agricultural and Rural Development Projects. FAO Nutrition in Agriculture Series (forthcoming). (Available in draft from CNSP)

Mason John, Habicht J-P, and Tabatabai H. (1982) Principles of Evaluation of On-going Programs, Cornell Nutritional Surveillance Program Working Paper Series No. 5, May 1982.

Mason J. B., Habicht J-P., Tabatabai H. and Valverde V. (1983) Nutritional Surveillance. October 1982.

National Academy of Sciences (1982). Nutritional Surveillance: A Synopsis, Washington, D. C. : NAS.

CNSP WORKING PAPER SERIES

Currently available:

1. Nutritional Surveillance: A synopsis. June 1981.
2. Nutritional Surveillance. August 1981.
3. Surveillance Summaries. January 1982.
4. Analytical Methods for Nutritional and Socio-economic Data. April 1983.
5. Principles for Evaluation of On-going Programs. May 1982.
6. An Outline Plan for Developing a Nutritional Surveillance System in New York State. May 1982.
7. Initial Assessment for Nutritional Surveillance in the Philippines. Forthcoming.
8. Preliminary Analysis of Nutritional and Agricultural Data, Eastern Province, Kenya. August 1982.
9. Static and Dynamic Poverty Indicators: An Indonesian Example. May 1983.
10. Reducing Malnutrition Through Rural Development: A test of Planning Methods in Haiti. June 1983.
12. Malnutrition by Socio-Economic Classification in Palawan, the Philippines. June 1983.
13. Changes in Nutritional Status in the Philippines from 1979 to 1980. Forthcoming.
14. National Health Information Systems and their use in Nutritional Surveillance. January 1983.
15. On Sampling Variance of the Sensitivity of an Indicator. January 1983.
16. Note on a Framework for Monitoring and Evaluation of UNICEF/WHO Nutrition Improvement Programmes. March 1983.
17. On Error in the Estimate of Malnutrition Due to Bias and Random Error in Anthropometry and Age. June 1983.
18. Age Misstatement for Children: A Problem for Interpreting Anthropometric Measures in Bangladesh. June 1983

Additional Documents Available from CNSP:

Report of Workshop on Social and Nutritional Surveillance in Eastern & Southern Africa. May 17-19, 1982.

Background Papers for Workshop on Social and Nutritional Surveillance in Eastern & Southern Africa. May 17-19, 1982.

8/83