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

Principles for Evaluation of On-going Programs
by
John Mason, Jean-Pierre Habicht & Hamid Tabatabai

This paper was originally prepared for a UNU workshop held at MIT in September 1981. It has been revised and submitted as Chapters I and II of "Methods for Evaluating the Impact of Food and Nutrition Programs," editors D.E. Sahn, R. Lockwood and N.S. Scrimshaw.

May 1982 **No. 5**

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 Science, New York State Colleges of Human Ecology and Agriculture and Life Sciences, Cornell University, Ithaca, New York 14853.

A report of research of the Cornell University Agricultural Experiment Station

TABLE OF CONTENTS

PART I: Basic Concepts for Design of Evaluations During Program Implementation... 1

Introduction..... 1
Costs of Evaluation..... 6
Experimental Basis for an Intervention..... 7
Purposes of Evaluation..... 7
Setting Program Objectives as a Basis for Evaluation.....12
Investigating Causality (link between program activities and outcome).....15
Confounding Variables and Gross vs Net Outcome.....19
Conventional Designs for Evaluation.....23
Levels of Analysis.....27
Definitions of Population Groups Involved.....31
Effect/cost.....35
Appropriate Indicators for Different Objectives.....36
Note on Sample Size.....46
Part I - References.....49

PART II: Stages of Evaluation of On-Going Programs----------51

Introduction----------51
Stage 1: Preliminary Tasks----------53
Stage 2: Evaluating the Plan of the Program----------55
Evaluating Implementation Objectives----------56
Evaluating Targetting Objectives----------57
Evaluating Outcome Objectives----------57
Stage 3: Evaluating Implementation----------63
Targetting and Delivery----------64
Improving Implementation----------64
Stage 4: Evaluating Gross Outcome----------65
Choosing Outcome Indicators----------65
Measuring Gross Outcome----------67
Stage 5: Evaluating Net Outcome----------69
Interrupted Time-Series Design----------71
Non-equivalent Control Group Design----------72
Stage 6: Data Presentation and Distribution of Results----------73
Stage 7: On-Going Evaluation----------75
References----------77

PART I

BASIC CONCEPTS FOR DESIGN OF EVALUATIONS DURING PROGRAM IMPLEMENTATION

Introduction

The original question we were asked to address was as follows: "What should be done to evaluate the effects of a program, at one point in time during the course of implementation (e.g. some years after the program began), under conditions where there are few baseline data, no control groups, and where results are needed quickly, which precludes time-series data collection?" A typical example was to be a food distribution program, such as supported by the World Food Program, although the techniques should be applicable to other sorts of programs. Such restrictive conditions do correspond to realities frequently encountered. Much of this and the next chapter accepts these constraints. We hope that evaluations may become earlier and better planned, so that some of the above constraints may be relaxed. Therefore we have also occasionally dealt with how to plan evaluations when one has the opportunity to go beyond the most basic levels.

Faced with the above questions and conditions, present practice is to collect data on program participants, and possibly nonparticipants, then to use statistical manipulations to investigate associations between program delivery and outcome variables. These methods tend to be expensive and may be difficult to apply in developing countries as routine procedures; moreover consideration of the questions which should be addressed before going into such methods reveals that they often turn out to be unnecessary. We therefore propose procedures which can be widely applicable to a range of programs in operation. These need to be distinguished from <u>experiments</u> designed to test whether a particular intervention <u>can</u> have an impact under pilot conditions.

Later we put forward a logical hierachy of fundamental questions which must be answered to establish whether a nutritional intervention <u>could</u> affect the performance, health and survival of individuals in a population (Table I.1). The procedures proposed here for the evaluation of on-going programs do not answer these questions, which are in the area of research - but they will give many of the answers which are of more immediate concern for a program manager, for an administrator or for a funder of programs. Whilst program evaluation of this type may not answer basic scientific questions, it must nevertheless be rigorous

TABLE I.1

IS A NUTRITION INTERVENTION JUSTIFIED?

Sequence of Questions to be Answered to Establish Whether a Nutrition Intervention Can in Principle Affect the Performance, Health and Survival of Individuals.

1. Is a deficit of food or specific nutrients causing disease, decreased performance or untimely death in individuals?

2. How detrimental is this deficit to individual performance, health or survival—in other words, what is the dose-response relationship?

3. Is it possible to decrease or eliminate the deficit (or its effects) in individuals?

4. How prevalent is the deficit (and its effects) in the population? Is the problem increasing or decreasing?

5. Is is possible to decrease or eliminate the deficit (or its effects) in the population?

6. What proportion of ill-health, decreased performance and untimely death in a population may be ascribed to this deficit now and in the future?

7. What are the expected benefits, costs and side-effects of interventions on the deficit (or its effects) in a large population, given the results of intervention trial studies in small populations? How long would the intervention need to run?

Table I.1 (Cont.)

8. What are the actual benefits, costs and side-effects of interventions undertaken in the large population on the deficit (or its effects)? Is the actual benefit worth the actual costs and the actual side-effects?

Source: Adapted from Habicht 1979.

and scientific in its logic. It must answer relevant questions, which must themselves be defined in advance. These questions are laid out in the section on "Purposes of Evaluation."

Evaluation terminology in this context (see for example Klein et al. 1979, Sahn and Pestronk 1981) differentiates "(basic) research", which is done to ascertain basic scientific facts independently of their applications to programs; from "evaluative research", which is done to assign a probability statement of causality to the relationships of an intervention and an outcome; and from "operational program evaluation", which ascertains whether a program is obtaining its objectives. Operational program evaluation is the subject of this and the next chapter. Operational program evaluation has also been divided into "summative evaluation" which examines the outcome of a program, and "formative evaluation" which monitors procedures and activities so as to improve the program by readjusting its services and their delivery. We refer to the former as outcome evaluation and to the latter as process evaluation. The document by Sahn and Pestronk (1981) provides a careful review of the theoretical literature in health and human services about program evaluation, and presents abstracts of many evaluations of nutrition intervention field programs - clearly differentiating evaluative research from operational program evaluation; other evaluative research studies of nutritional programs are reviewed by Habicht and Butz (1979).

There is in practice no clear distinction between research and evaluation in the methods used. There is, however, a major difference in the kind of questions being addressed, and the appropriate parsimonious combination of methods to address that question. It is a grievous mistake to call research "more scientific" than evaluation. Good science is fitting the appropriate methods to seek an answer. A good evaluation is as scientific in this sense as is good research.

Unfortunately most works on operational program evaluation in nutrition have been misdirected in emphasizing the researcher's concern to substantiate the probability of causality, a research question, while neglecting the other important questions which managers, administrators and funders ask. This misdirected research methodology therefore often neglects formative or process evaluation, which is a sine qua non of a successful evaluation. This error is often compounded by emphasizing measurement of outcome variables, which is probably the least of the weaknesses in past nutrition intervention evaluations.

Many of the researchers tools can be used to answer questions relevant to evaluation. In the next chapter, a step-wise procedure is proposed in which the

easier and cheaper investigations are done first to examine whether a program could be expected to have an effect based on its design; whether it has been implemented in such a way that the effect could have been achieved in practice; whether there is reasonable evidence to expect that the effect has actually been achieved; and finally to investigate in more detail whether the effect is due to the program.

J.D. Wray in preparing an actual evaluation of nutritional activities out of a rural health center in Colombia claimed that the reasons why evaluation of nutrition intervention activities were not done as often as they should be was because of a lack of motivation, which included a failure to realize the importance of evaluation, in part because of lack of experience in using evaluation results, and a lack of courage to specify objectives and put the program to the test of an evaluation of the success in meeting these objectives (Wray 1970, 1973). We agree, but also feel that the reason evaluations of programs are seldom done is because of a lack of understanding of the purposes of an evaluation, and therefore there is a poor fit between the methods proposed for evaluation and their implicit purposes.

This chapter is intended to introduce a number of basic concepts needed for designing evaluations appropriate to the decisions to be taken based on their findings. A first consideration in practice is the **resources available** for evaluation (see 'costs of evaluations'). Programs should have been based on evidence from basic and evaluative research which shows that they can have the desired effect - although not part of evaluation, this fundamental concept is touched on in **experimental basis for an intervention.** The **purposes of evaluation,** for different audiences, are then related to the available tools and designs (see 'purposes of evaluation'). **Setting program objectives**, or investigating these, is logically the next step after the evaluation objectives have been decided, and some principles are laid out in "setting program objectives as a basis for evaluation".

A crucial analytical question concerns interpretation of **causality:** whether a measured outcome can be ascribed to program activities. The theoretical sequencing of issues and questions is given in some detail, which sets the context

for the next sections on **confounding, conventional evaluation designs, levels of analysis** and **definitions of population groups involved.** A brief review of the issues related to **effect per cost** is then given. Finally, a section sets out **appropriate indicators for different objectives,** by program type, with a note on **sample sizes.**

These concepts are intended to provide the basis for the practical stages in evaluation given in Part II.

Costs of Evaluation

Useful evaluations entail activities which apparently are not necessary to manage the logistics of a program, and hence incur additional costs. It has been said (Riecken 1979, p. 365) that "...my experience with evaluation is that there are few bargains, and usually you get no more than you pay for" and that "...it is too simplistic to say that complexity is hard to understand, takes a long time, and costs a lot of money, while simplicity is easy to understand, done quickly, and is cheaper. When an evaluation is cheap and quick, it is often also not very good." Without entering into this debate, we can nevertheless learn something from the actual levels of expenditure on evaluation. On the one hand, the World Bank review of monitoring and evaluation of projects in East Africa in 1979 (Deboeck and Kinsey 1980), for example, gives a figure of $12.8 million for monitoring and evaluation of 28 projects averaging $460,000 each; or somewhere around 0.5 to 5% of total project costs. The U.S. Government estimates that 1% should be used to evaluate programs in health. On the other hand, Kielmann et al. (1980 p. 33) reckons that "...it should be neither uncommon nor unreasonable to budget 20% to 40% of total project costs for analysis of project generated data and project evaluation". Evidently the scale of expenditure of the project itself has some influence on these calculations. Similarly the purpose of the evaluation of the project is important - whilst relatively high expenditures may be essential for pilot or experimental projects, it seems unlikely that more than perhaps 5% of project costs would be made available for routine evaluations of large-scale service projects. Such expenditures would often be sufficient to allow useful evaluation, if it is recognized that causality may not be established. Evaluation should be able to go as far as assessing what we refer to as the "adequacy" of the program. This adequacy evaluation is likely to be more effective when a means of evaluation is built into the program; this is discussed in Mason et al. (1982, Chapter IV).

Experimental basis for an intervention

The evaluations considered here do not attempt to answer questions concerning the scientific basis for interventions. These are best tackled through experimental research. However, unless these have been addressed in designing the intervention, the evaluation may be pointless. The logic for determining whether an intervention can in principle have an effect - its scientific basis - is introduced here as shown in Table I.1, because it must be borne in mind whenever an evaluation is considered. Questions 1 to 4 correspond to basic or experimental research, and 5 to 8 to evaluative research (see Sahn and Pestronk 1981). These are prerequisites for the operational program evaluation which is the subject of these chapters, but are not part of it.

It is crucially important to emphasize that none of the questions in Table I.1 are the object of operational program evaluation. They are research not evaluation questions. However the answers to these research questions are necessary to evaluate whether a program is likely to have an effect. This has been referred to as "relevance" (Beghin 1980). Relevance essentially means "is the intervention addressing the problem, or at least one cause that is accessible to intervention?". For example, poverty, an unhygienic environment, and contaminated food or water may all be causes, at some level, of child deaths from diarrhea. Interventions at any of these points could be justified on grounds of relevance. On the other hand, it might be held that inappropriate weaning practices cause children to eat a diet inadequate in protein; but further study might show that protein intakes are not the limiting factor; hence a supplementary food program aimed specifically at increasing protein intake would be irrelevant.

Purposes of Evaluation

The question **"How do we tell if a program has an effect?"** is incomplete without knowing why one needs to know. Common reasons are - "to know whether the effect is adequate, in order to...":

- decide whether to continue the existing program or not
- redesign the program if necessary
- decide whether to do similar programs elsewhere

Those involved in the program often have different expectations about the purpose and results of an evaluation. It is important that the decisions ("in order

TABLE I.2

BASIC ISSUES IN DESIGNING AN EVALUATION

Evaluation for whom?

- o Managers (M)
- o Administrators and Funders (A)
- o Researchers (R)

To determine whether the intervention

- o is performing as expected (M, A)
- o worth continuing (A)
- o should be extended (A)
- o is causally linked to improved nutrition (R)

to") which the evaluation will affect be clearly understood and agreed upon. The evaluators must then tailor not only the design of the evaluation to the purpose of the evaluation, but also the presentation of the results to that purpose. For instance, results presented as if the purpose is to decide on the continuation or termination of a program are inappropriate if the purpose of the evaluation is to improve the program. Evaluation cannot be seen in isolation from who asks the question. The situation is not so much that the principles and practice of evaluation of ongoing programs are unsatisfactory, but that the whole decision-making process in nutrition and food aid programs needs improvement.

The purposes and issues addressed in evaluation depend on who is asking the questions: these can be program managers, administrators, funders, or those concerned (often researchers or scientists) with basic mechanisms of cause and effect. The different interests are outlined in Table I.2.

The sequence begins with considering whether the program is performing adequately and can progress to seeking to ascribe causality between the intervention and the outcome. This sequence approximates to the changing concerns of project management, administrators, and to the researchers concern with causality; however, causality, if it can be shown, is also of important interest in all aspects of management, program design, and policy. The trouble is that causality is difficult and expensive to establish, and the more that certainty on causality can be dispensed with, the easier the evaluation becomes. For example, project management can often in fact get by with the knowledge that the beneficiaries are improving, even if they cannot be sure this is due to the program.

Part of the information needed to address questions such as those given in Table I.2 can be obtained from evaluation of the project design and from **process** data. Moreover these data can be used to screen out those projects which are unlikely to be having any important effect on outcome, which it is then not worth trying to evaluate further. This is set out in subsequent sections. There are other decisions in establishing purposes of an evaluation that center on the degree of certainty required in linking **outcome** to program delivery, which need to be explained in more detail at this stage.

TABLE I.3

APPROPRIATE DATA COLLECTION AND ANALYSIS FOR DIFFERENT DECISIONS

Confidence needed increases →

Data & Analysis Needs (Difficulty Increases ↓)	Decisions	(1) Management	(2) Continue Funding	(3) Replication in similar conditions	(4) Replication in different conditions	(5) Basic research, causality
(a)	Process data & outcome for participants only	+	+	+	+	+
(b)	Ad hoc surveys	(+)	(+)	+	+	(+)
(c)	Advanced stat. analysis		+	+	+	+
(d)	Some kind of control group(s)			+	+	+
(e)	Before/after data				(+)	+
(f)	Highly standardized measurements				(+)	+
(g)	Randomized intervention				(+)	+
(h)	Double-blind trials					+?

(+) Means occasionally

Different purposes of evaluation demand varying degrees of plausibility or certainty for the conclusions reached from the evaluation. The purposes, in the order of increasing need for higher levels of certainty (and elaborating from Table I.2) are:

(1) improvement in program management

(2) continuation of funding

(3) replication of the program in similar conditions

(4) replication of the program in dissimilar conditions

(5) finding basic research results about cause-effect relationships.

The methodological and data requirements of responding to the differing needs of these purposes for certainty and plausibility entail, in order of increasing expense and difficulty:

(a) collection of data on process and outcome for participants only (program data)

(b) collection of data through ad hoc surveys

(c) advanced statistical analysis

(d) control group(s) of some kind

(e) collection of before-after data

(f) highly standardized measurements

(g) randomized intervention

(h) double blind research designs (blind intervention and blind assessment)

It would be useful to consider how these two lists could be matched. Each item in the first list is taken up individually; this discussion is summarized in Table I.3.

(1) Evaluation for program management seeks to determine whether program services are being delivered as planned to the intended target groups and whether the (gross) outcome is acceptable. The objective of this "adequacy evaluation" is to reveal the possibilities for improvement in program management. It does this by relying on program data relating to process and participants (method (a)). It may, on occasion, require survey data as well, i.e. (b).

(2) The decision as to whether to continue funding of a particular program often requires adding advanced statistical analysis (c) to the requirements for adequacy evaluation, i.e. (a), and possibly survey work (b).

(3) Replication of the program in similar conditions usually requires data on some form of control groups (d), and/or surveys (b), in addition to the requirements for (2).

(4) Replication in dissimilar conditions entails, at the minimum, both control groups (d) and survey data (b), as well as advanced analysis (c). Sometimes it may be preferable to use a quasi-experimental design (see section on Design below) e.g. to add before - after data (e), to standardize measurements carefully (f) and/or to use a randomized design (g).

(5) Basic research involves most of these requirements together, and sometimes it may even be possible to employ the ideal research design, the double-blind randomized trial (h).

The confidence with which the conclusions in each of the above cases is reached can be considerably improved with strong theory relying on good scientific evidence from elsewhere. If such a theory exists, the methodological tools and data in each case may be sufficient to provide the level of confidence required by the one immediately following it.

Setting program objectives as a basis for evaluation

Although it may often not be feasible to determine absolute values of change in relation to program activities, it is essential to decide in advance whether certain differences are adequate. These provide objectives against which the program can be evaluated. For the manager's questions (e.g is the program performing as expected? - see Table I.2) and usually for the funder-administrator's questions (e.g. is the program worth continuing, extending, etc. - see Table I.2) the comparison is between the procedures and activities of the program and some preset standards, generally set out in the program work-plan. The first prerequisite is, therefore, that the essential activities be stated in objectively measurable units. This is possible even for such an amorphous exercise as curative primary health care (Habicht 1979). Actual performance relative to these standards is ascertained through **process evaluation.**

A requirement for **outcome evaluation** is to establish outcome objectives prior to assessment. These objectives of the program must be explicitly formulated as an acceptable difference from a standard, or as a minimum improvement from some baseline. These quantitative standards of achievement

should correspond to the implicit objectives of the program, and should be understood and agreed upon by those who must use the results of the evaluation. Experience shows that while one is working with those involved with a program to make the stated and implicit objectives more explicit, one will often find unexpected objectives, some of which are even contradictory. This is why a consensus about the program objectives is one of the necessary first steps to an evaluation.

Standards of procedure, activity performance and outcome are essential to adequate management, but no effort is visible yet in the nutrition program literature as to how this should be done so that the standards are useful. Now that the results of a number of carefully measured and analyzed pilot nutrition and health programs are becoming available (e.g. Habicht and Butz 1979 Drake et al. 1980; Gwatkin et al. 1981), it should be possible to suggest feasible outcome objectives for adequately funded and managed programs. The results cited in these reviews tend to indicate that approximately a five to fifteen percentage point reduction in malnutrition prevalence can be achieved by certain types of programs.

Almost inevitably program objectives change as a program evolves. However, changing definition of objectives <u>during</u> evaluation of <u>single</u> projects should be avoided because it is rare that the design of the evaluation can deal with new objectives. For example, in a recent review of supplementary feeding programs (Beaton and Ghassemi 1979), because the supposed objectives of improving child nutrition were seldom reached, there was a discussion as to whether the more important effects of these programs were in terms of income distribution. However, no comparison was made with quite different programs that might be more efficient in changing income distribution. While this may be a reasonable question in general, changing objectives for an individual program requires more fundamental decisions.

Once the underlying outcome is identified in a conceptual sense comes the step of identifying the measurable variables which are related to the outcome of concern. The major portion of this book discusses that step which relates a conceptual outcome to a measured variable. Many of these variables are discussed later and summarized in Table I.10 according to the conceptual outcome as it relates to a specific kind of intervention and as to the resources

required to take the measurements. Subsequent chapters in this book develop the relationship between the conceptual outcome and the measurements more fully.

Finally, an a priori calculation of expected costs with relation to expected effects should be made as part of setting objectives. Here again, there is a dearth of empirical data, although again some crucially important results are now becoming available (e.g. Gwatkin et al. 1981). There is an urgent need for more research in this area to provide a more rational basis for planning and evaluations.

The statistical test used to judge the reality of a measured difference (either between treatment and control groups, or between treatment and a standard) results in a statement that most of the time (which is usually specified as 95 or 99 percent of the time) such a measured difference will be found if the true difference is not smaller than some quantity. In designing an evaluation one must further state how often one is willing to miss identifying a true difference of more than a specified magnitude. This statement refers to power analyses (see e.g. Cohen 1969). These steps of specifying procedural and impact objectives, translating those objectives into conceptual outcomes, transforming those outcomes into measurable variables, specifying the minimum or maximum acceptable difference in that variable and doing the power analysis are prerequisites for any quantified evaluation. If the research referred to in Table I.1 has been done, which should in principle always be the case before a program is implemented, none of these steps should be too difficult or time consuming. The sad fact is that the research giving scientific justification to a program is often so lacking that these steps are impossible. Experiments in the precise situation of the proposed program may not always be needed (or possible); but a marshalling of the evidence, from experiments elsewhere, from previous evaluations, and from scientific knowledge, as a basis for designing a relevant program is all too seldom done. Wishful thinking then takes over. This leads to a program for which there is no solid reason to expect a measurable outcome. There is then no need to go further in the evaluation — the basis for operational program evaluation is lacking. There is no way that an evaluation of an ongoing project can deliver any useful results in these circumstances.

Investigating causality (link between program activities and outcome)

A summary of the logical sequence of questions in evaluating whether a measured outcome is plausibly caused by program activities is given in Table I.4.

The basic question is: "Is there a statistical association between putative cause (the intervention) and outcome or effect (e.g. improving nutritional status)?"

Association itself requires comparison of measured outcome in at least two groups which receive different intensities of program intervention. Seeking associations is only useful when one needs some level of certainty that the program is causally related to the outcome. The association itself may be shown by comparison of two groups (e.g. control and treatment), by showing correlations between different levels of program delivery and outcome, by estimating regression coefficients between program activities and outcome statistically controlling for other influences, and so on. Controlling for other than program influences on the outcome, which mimick a program effect, is called "controlling for **confounding** influences". The questions to be asked about a program and its evaluation are very different depending on the results of the statistical tests of association. Although these tests are performed after data collection they must be foreseen to insure that the right data are collected. The first question after data collection and analysis are completed will be:

"Was there a statistical association between the program intervention and the outcome?"

If **no association between the intervention and outcome** is found, the next questions are as follows:

1. Are the probability statements themselves correct in stating that there is not a high probability of association between outcome and intervention: were the right statistical tests used given the objective of the program?
2. Was the intervention relevant; was it adequately applied? Check program design and process evaluation.
3. Were the measures used of improved nutritional outcome relevant to the program objectives; there is no use using indicators which have not been shown to be responsive to the intervention (see Habicht and Butz 1979; and Table I.9 below).

TABLE I.4

EVALUATING OUTCOME

- Is the outcome adequate? (M, A)
- Is there a statistical association between intervention and outcome? (A, R)
- Is the outcome due to the intervention? (A, R)

[1] A, M, R as Table II.2

<u>If there is no statistical association or the outcome is inadequate;</u>

- Were the statistical methods used correct?
- Was the intervention relevant and adequate?
- Were the measurements of outcome valid and reliable?
- Were the recipients likely to benefit?
- Was the sample size adequate?
- Was there negative confounding?

<u>If there is statistical association and outcome is adequate:</u>

- Was the association likely to be causal?

  Discard confounding - see Table II.5 on internal validity
- What direction was the causality?
- What mechanisms linked the intervention to the outcome?
- Can the findings be extrapolated to the population as a whole or to other populations? (i.e. external validity)
- What was cost of the effect and of the marginal effect?

4. Were the recipients likely to show a benefit? This question is a different way of posing questions 2 and 3, but should be asked explicitly.
5. Given yes to 1 to 4, was the sample size adequate to reveal a meaningful minimum association? In whom (needy, targetted, participant population, whole population - see section on "definitions of population involved")?

This can be checked as follows:

(a) What size of effect is expected under these conditions?

(b) What size are errors of unreliability due to:

- Measurement imprecision
- Undependability due to random influences on outcome

These errors due to unreliability can be reduced by:

- Exclusion by design;
- Stratification before treatment and exclusion by analysis (e.g. matching);
- Measurement and exclusion by analyses (e.g. multivariable analysis)

(c) Recalculate necessary sample size given (a) and (b) and desired probability of type I and II errors.

(4) Was there a countervailing influence which cancelled out the association: e.g. treatment contamination, negative confounding, unintended treatment of controls (especially if dose-response only large at lower levels), those not receiving the program compensate for its lack (if done from bravadura it is called a "John Henry effect" (Saretsky 1972)).

If **an association between the intervention and outcome is found** the next questions are as follows:

(1) Are the statistical probability statements themselves incorrect due to:

- Fishing (multiple significant tests)
- Wrong assumptions about the appropriateness of chosen statistical tests for the data.

(2) What are the boundaries of the magnitude (not the statistical significance) of the association? E.g. regression coefficients

corrected and uncorrected for unreliability of independent variables. Is the putative effect adequate? Adequacy depends on program objectives.

(3) Was the association causal?

Discard confounding:

(a) Inherent in comparison groups:

Exclude by design:

- randomization of treatment*
- measurement of confounding variables and exclusion by analysis. E.g. drop- outs, covariance analysis, etc.

(b) Inherent in treatment (e.g. treatment carries unidentified true cause, those being watched performed better)

- exclude Westinghouse effect by blind treatment -placebo*
- exclude unidentified true cause by better characterization of treatment and counter experiments with contaminants

(c) Inherent in outcome measures

- standardization alone will not fully exclude this confounding
- exclude by blind assessment*

(4) Once causality is established, what was direction of causality?

Introduce time sequence:

(a) by intervention

(b) by natural experiment (e.g. as in quasi- experiments) taking advantage of and measuring changes of independent variable over time.

(5) What are the mechanisms of the causal relationship?

(6) What is the cost of the effect? What are the marginal costs?

(7) Can one extrapolate these findings to the future or to other situations?

This list of questions is reducing in its logic, but provides in fact little guidance in designing an evaluation in real life, without better definition of which questions are relevant to the objective of the evaluation: to provide guidance to the program manager, or the funder, or the researcher.

The list of questions following the conclusion that there is not association is useful also if one finds that an evaluation is not achieving some level of preset outcome. These questions are therefore universally useful if the program appears to be producing inadequate results.

The list of questions following the conclusion that there is an association, is irrelevant to an evaluation of a measured outcome relative to a preset standard. Some of the questions in items 3 and 4 are so difficult to answer perfectly in the context of a realistic program evaluation that one must examine carefully whether they are really necessary or even useful for the evaluation. The perfect solution to confounding is to design an intervention in such a way that all influences on outcome other than that of the program are randomly distributed among comparison groups. This then permits one to state that a probability of association is a probability of causation. Any other statement about causality is one of plausibility to which one cannot ascribe a probability statement. This is important because it means that changing criteria of statistical probability testing (i.e. increasing $P < x$ from $P < x$ 0.05 to $P < 0.1$ or $P < 0.25$) is completely irrelevant to designing feasible program evaluations, because it does not make an evaluation design which delivers plausible statements about causation into one which delivers a probability statement about causation.

The random distribution of all (confounding) influences on outcome other than that of the program requires designs which are impossible in evaluation and these are starred (*) in the list above, and will not be discussed further.

Evaluation must therefore deal with confounding influences in a different fashion to establish the validity of the results, and thus be able to make a plausible statement about the program being the cause of the improvement.

In the following sections, some of these concepts are elaborated to provide background for the procedures discussed in Part II.

Confounding Variables and Gross vs Net Outcome.

The validity of a research design is a measure of the degree to which its conclusions reflect real phenomena. There are four types of validities: construct validity, conclusion validity, internal validity and external validity (see Judd and Kenny 1981). Construct validity refers to the extent to which the nutritional factor of interest has been successfully operationalized, i.e. the extent to which the particular set of outcome indicators, treatments, population

TABLE I.5

Major Causes of Confounding (threats to internal validity)

Cause of Confounding	Example
1. SELECTION: When the assignment of subjects to treatment and comparison groups is not random, the groups may differ systematically in some characteristic(s) associated with the outcome variable. Selection bias is therefore likely to be present. Self selection is a common source of this type of bias.	Mothers who choose to participate in a program to reduce the incidence of low birthweight may tend to be more educated, richer and more motivated than those who do not. These factors influence the outcome and compete with the program as an explanation for an observed reduction in the incidence of low birthweight.
2. MATURATION: Human subjects mature overtime and this process may cause changes in the outcome variable irrespective of program effects.	The nutritional status of 6-24 month old children is often worse than that of older preschoolers. If the average age of participants in a nutrition program increases from, say, 18 months to 36 months, observed improvement may be due to maturation and would have occurred without the program.
3. HISTORY: When a program is in effect, many other events may intervene and influence the outcome variable. When these historical events have different impacts on the treatment and comparison groups, they confound the program effects.	A supplementary feeding program is introduced in one of two otherwise equivalent areas. If food prices rise at different rates in the two areas the observed differences in nutritional status may not be attributed solely to the feeding program. Differential food price rises may have influenced the outcome.
4. INSTRUMENTATION: This threat arises from changes in how measurements are made or what is measured or from measurement errors due to, for example, instrument decay.	The height for age of preschool children is often compared across age groups. Since it is usually the lengths of 0-2 year old infants that are measured as opposed to the heights of children older than 2, such comparisons may be biased.

TABLE I.5

Major Causes of Confounding
(threats to internal validity)

Cause of Confounding	Example
5. REGRESSION ARTIFACT: If subjects are chosen on the basis of exhibiting extreme value on some variable (e.g. wasting) there may be an improvement overtime without any intervention. This tendency is called regression toward the mean. The solutions are either to observe the effect on the whole population, or to make comparisons within the selected extreme group (e.g. the malnourished).	Nutrition programs instituted for the malnourished may show an an improvement in that some participants may no longer be malnourished at the end of the program. Part of this improvement may not be due to the program since some subjects would have improved anyway.
6. EXPERIMENTAL MORTALITY: Some subjects may drop out of a program in the course of its implementation. If these subjects have different characteristics than those that remain, any before/after effect shown may be confounded by differences in the populations before and after the program.	A food-for-work program may not lead to an improvement in the nutritional status of a community even if it has in fact been effective. This could happen if enough of the participants who improve leave the community in search of jobs elsewhere. The observed change here underestimates the impact of the program.

and settings truly reflect the desired theoretical nutritional construct. The operationalization of the notion of nutritional status of a community, for example, involves the use of standardized anthropometric indicators. How successfully the chosen outcome indicators in fact reflect the underlying construct of interest defines the degree of success of the operationalization.

Conclusion validity refers to the ability of a research design to allow an investigator to detect outcome effects if they exist. This definition emphasizes the importance of trying to avoid concluding that no effect exists when in fact it does ('type II error'); as opposed to wrongly concluding that an effect exists when in fact it does not ('type I error'). Committing a type II error however can be quite costly in that large scale programs that have been instituted and are in fact effective are often concluded to have insignificant effects because of inappropriate balance between the probabilities of these two types of errors and hence the inability of the evaluation design to detect their effect.

Internal validity of a design refers to the extent to which the detected outcome changes can be attributed to the intervention or treatment rather than to other causes. Unless the internal validity of a design is high , the finding that a particular relationship is causal will not be particularly convincing. Some of the major threats to the internal validity (i.e. confounding factors) are summarized in Table I.5. The primary reason for the proper choice of a comparison group and for statistical adjustment techniques is to control for these threats as best as possible, but without randomized allocation.

External validity pertains to the generalization of the conclusions to other populations, settings, etc. Statistical adjustment is generally irrelevant here although the subject matter itself is a major determining factor of the level of external validity.

The expression "gross outcome" is used to refer to the measured change in outcome variables without allowing for change that would anyway have occurred as a result of other factors - i.e. for confounding variables. "Net outcome," which is synonymous with "impact", allows for these.

Conventional designs for evaluation.

Conventional designs for evaluation are commonly considered in three groups: non-experimental designs, quasi-experimental designs, and true experimental designs. There are many standard works on this subject (Cook and Campbell 1979; Poister 1978; and Judd and Kenny 1981). Selected designs considered here are summarized in Table I.6.

The use of **non-experimental designs** is widespread in social program evaluations. They consist of observations of outcome with no attempt to control for confounding variables. This lack of control poses a number of threats to internal validity and hence the conclusions based on such designs are usually only descriptive and cannot of themselves say anything about the effectiveness of a program. They are, however, very useful, and all that is necessary for adequacy evaluation to know if the program outcome is reaching preset standards.

In diagramming different designs, as given below and in table II.6, X stands for "treatment" (program) and O for an observation of outcome variables in a population or a sample of the population (unless otherwise indicated), and time moves from left to right.

The simplest case of a non-experimental design, the so-called "one-shot case study" can be depicted as: X O. In this design the observations are taken after the program treatment has been completed or has been in operation long enough to have an effect. An example of the use of a "one-shot case-study" would be to measure outcome variables - say nutritional status - in program participants only, with no measured variation in program delivery. This can lead to conclusions as to whether or not nutritional status is now "adequate", either with respect to program objectives or some other criteria. It <u>cannot</u> provide information as to whether this is due to the program, nor (hence) whether nutritional status would deteriorate without the program. On a one-time basis this design may be useful to program management but is of limited use to those who must decide about continuing the program since it does not define what effects the program itself has had. There are two somewhat stronger non-experimental designs that allow for some kind of comparison: the one-group pretest/posttest design and the static group comparison.

The one-group pretest/posttest design: O X O allows for the comparison of the preprogram situation with that postprogram. It does not take account of changes that might have anyway occurred. If the measured change is substantial, particularly if compared with an expected no change otherwise, this

TABLE I.6

CONVENTIONAL EVALUATION DESIGNS

Notation:

O = Observation of outcome for 1 to 3, of both outcome and confounding variables for 4 to 6
x = treatment

Time →

DESIGN	REFERRED TO AS	ANALYSIS	DELIVERS
1. XO	One-shot case study	None	Adequacy
2. OXO	One-group pre-test/ post-test	Compare before/after	Adequacy
3. Group 1 XO Group 2 O	Static group comparison	Compare groups	Adequacy
4. X (Varies) O	Correlational	(a) Compare sub-groups (b) Correlate treatment levels with outcome, controlling for those confounding variables measured which are not themselves highly correlated with treatment	Adequacy, some inference on net outcome
5. Group 1 OXO Group 2 O O	Non-equivalent control group design	Compare groups with statistical control for confounding	More plausible inferences on net outcome
6. OOO X OOO	Interrupted time series	Before/after; time-series	" " "

design might give some indication of program effect. It is particularly important here not to only consider the extreme cases, such as the severely malnourished, because they might often be expected to show some remission in the normal course of events. This "regression to the mean" is a well-known confounding phenomenon, (see e.g. Davis 1976, Furby 1973). Here it is important to consider the total population from whom those treated are selected, and a decreased prevalence in the population as a whole is more convincing than a decreased prevalence in the group screened out as needy at the beginning, unless the expected rate of recovery is known.

This eventuality and others may sometimes be controlled for in the static group comparison design:

Group 1	X	O
Group 2		O

This design shows a program instituted at time (X) followed by measurement (0) in both the program and another population at a later time. Following the above example, groups 1 and 2 should ideally be equally malnourished before treatment to allow any comparison. By definition in this design this is not known, but may possibly be inferred from other data. Including pre-program data turns this into a more powerful design - see no. 5 in Table I.6.

A more effective, but still limited non-experimental design is the correlational study carried out on a cross-sectional basis (i.e. at one time only). This design relies on a large enough cross-section of the participant population, who have been subjected to the program with varying intensity, and depends on collecting data on potential confounding variables which may independently influence the outcome, as well as the outcome variables themselves. If the number of observations is large enough and if the observations on each variable are significantly diverse, then it may be possible to use statistical techniques of control to do what the experimental conditions preclude, namely, to allow for the influence of the confounding variables chosen (see no. 4 in Table I.6). Because of the problems associated with this type of design (e.g. see Poister, 1978, pp. 271-2), its primary usefulness is not in establishing causal relationships but rather in suggesting hypotheses that should later be more carefully examined. It can increase the plausibility of inferences on causality, but requires more advanced analytical techniques.

An example of correlational studies would be to collect data on a cross-section of the population, measuring a range of factors possibly affecting

outcome, including data on program delivery. The analysis would then examine the degree of correlation of outcome with program delivery taking account of other factors. In its usual form analysis involves multiple regression techniques requiring extensive computer facilities; multiple group comparisons, which are more easily calculated, may give similar answers in some circumstances - see next section.

When feasible, correlational analyses will deliver the most plausibility under the restrictive conditions of no comparison groups and no pre-program data, as originally put to us (see first paragraph of this chapter). However, they are demanding in terms of data and analysis, and whenever possible more powerful designs should be used, to include some form of comparison groups and/or time-series data: for example the quasi-experimental designs discussed below.

At the other end of the spectrum from non-experimental designs are true experimental designs, in which subjects are randomly assigned to the treatment and control groups. These are not appropriate to most program evaluations, because their need for random assignment can deny benefits to the needy only for the sake of evaluation. They are not discussed further here. They are, however, the only designs that will allow a probability statement about the causality of the effect of the program.

A compromise to the true experimental design that has been widely used has sometimes been referred to as **quasi-experimental** design. These try to strike a balance between the stringent refinements of true experiments and the weaknesses of the non-experimental designs. Their principal difference with the true experimental designs is that randomized assignment of subjects to treatment and control groups is not practiced. There are many examples of quasi-experimental designs (for a comprehensive review see Cook and Campbell 1979). Two of the basic and most frequently employed examples, which have been put forward in the context of agricultural and rural development projects by the World Bank (Casley and Lury 1982), are the <u>non-equivalent control group design</u> and the <u>interrupted time-series design.</u>

The nonequivalent control group design can be diagrammed as:

Group 1	0	X	0
Group 2	0		0

It applies to situations where random assignment is not feasible, but some selection of comparison groups is possible. The treatment and comparison groups are chosen before the program begins in such a way that they are as similar as possible with respect to variables judged to be important in determining outcome. This matching is expected to control for some threats to the internal validity of the evaluation, but it has little power to control for the selection bias. Despite an attempt to find similar groups, it is still not possible to be sure that the presumed impact of the program does not in fact arise from the differences that may exist among the two groups of subjects. The analysis in this case has similarities to that for a correlational analysis; preferably multivariable techniques are used to control for measured differences due to imperfect matching.

The interrupted time series design can be shown as:

One group only 0 0 0 0 X 0 0 0 0

This design seeks to attribute an impact to a program by making a series of observations over time both before and after the treatment. If the treatment is of relatively short duration and if it makes its impact relatively quickly, this may be a suitable design to aim for. Many public programs however are in effect for a long time and their impact too is only perceptible over a comparable period of time. As pointed out by Casley and Lury (1982) when there is a "sharp break" at some point due to treatment, this design may be of particular use.

The choice of these techniques involves a trade-off between the difficulties of data collection (a) on comparison groups and (b) over time, with the plausibility of the causal inference drawn. They also depend to some extent on the analytical capacity available, as shown in the third column of Table I.6 and discussed in the next section.

Levels of Analysis.

Decisions on levels of analysis to be used are important because:

(a) the necessary skills for complex analyses may be lacking in developing countries;

(b) interpretation can be improved by advanced analyses, at least in terms of plausibility of conclusions;

(c) time and cost are in reality related heavily to the extent of analysis.

More advanced analyses have on occasions modified conclusions and often this has been clarifying. In some cases the clarification of further analysis avoided wrong conclusions which could in fact have been detected with commonsense looking at the data. For example, introducing socio-economic status into an analysis of the apparent effect of the Narangwal experiment actually reversed the apparent direction of effect of the program (see Chernichovsky 1979). Probably, however, this conclusion could still have been reached simply by dividing the sample into a two or more socio-economic groups. On the other hand, unresolved differences in conclusions between investigators using different analytical techniques have also been known to result, because the assumptions underlying the analyses were not the same. Often the investigators did not realize this discrepancy themselves. This may warn against too much reliance on too sophisticated techniques particularly in developing countries. This is particularly so since advanced analyses may be not widely feasible, and certainly efforts should be made to seek the simplest analytical procedures - and this starts with the design of the evaluation.

We distinguish between "basic" and "advanced" analyses. Basic analysis refers to: categorical data analysis - for comparison of frequencies (e.g. prevalences) between groups; correlation analysis, for investigating the degree of association between two variables (e.g. whether prevalence of malnutrition is correlated with a possible determinant); and analysis of variance, used to determine whether differences exist between mean values of indicators for a number of groups. The methods of advanced analysis reckoned to be most suitable for the problems we are interested in are the methods of multivariable analyses (e.g. ordinary least squares regression analysis, discriminant analysis, logit analysis, probit analysis, etc.), for investigating associations between outcome and a number of possible determinants, in this case obviously including program delivery.

In deciding the overall plan of an evaluation, a balance needs to be struck between design, extent of data collection, level of analysis and plausibility or certainty required of the conclusions. To some extent with good design, less sophisticated analysis is required; for example designs with adequate control groups, or before-after data, may require less investment in both data collection and analysis than an uncontrolled (by design) post-program correlational analysis. The appropriate analyses by design are indicated in the third column of Table I.6.

When the capacity for advanced statistical analysis is not available - which may frequently be the case particularly in poor countries - a lot can still be achieved by commonsense treatment of the data and by comparison of suitably defined groups. Indeed, even when more advanced techniques are used, it is important to be clear conceptually which groups are being considered. For example, very often socio-economic status and/or sanitary conditions are a primary determinant of differences in outcome of variables such as nutritional status or health: they can confound conclusions on program effects. Both these can be measured: socio-economic status for example by income, quality of housing etc. Analyses then are done by suitable groupings.

If program delivery varies - even if there is no "no program" group as such - tabulation of results such as in Table I.7a can be informative and valid. The interpretation of different options could be as follows. Example 1, in which the only group with poor nutritional status is that with low socio-economic status and poor program delivery, tends to indicate that the program is having an effect. The conclusion from this is possibly that delivery should be improved to the bad socio-economic group. Example 2 indicates that most of the difference in nutritional status is being acounted for by socio-economic factors, and that more detailed examination of whether the program can have an effect is needed. Example 3 indicates that the program is related to most of the differences in nutritional status. It also indicates that more efficient delivery is required because those not receiving the program could benefit from it. Additional confounding variables - such as sanitation - could be added to such a table although numbers per cell would decrease. Moreover information maybe lost by categorizing socio-economic status in this way, if it can be measured as a continuous variable. To combine several variables and make the most use of the available information, multiple regression techniques are often applied. For evaluation, the outcome (nutritional status) is the dependent variable, and program delivery is treated as one independent variable along with other determinants (confounding variables) such as in this example socio-economic status and sanitation. The purpose is then to examine the significance (in a statistical sense) and importance (of the magnitude of the effect) when other determinants are allowed for. It must be emphasized, however, that when the substantial computing power required for multiple regression is not available, tabulations by group, for example, as in Table I.7a can still give important results.

TABLE I.7

EXAMPLE OF COMPARING OUTCOME WITH PROGRAM DELIVERY AND SOCIO-ECONOMIC STATUS

A. **PROGRAM DELIVERY**

	HIGH		LOW	
Example	SES HIGH	SES LOW	SES HIGH	SES LOW
1	+	+	+	-
2	+	-	+	-
3	+	+	-	-

B. **PROGRAM DELIVERY NOT VARYING**

Example	SES HIGH	SFS LOW
4	+	-
5	+	+
6	-	-
7	-	+

+ Means satisfactory outcome indicator values - e.g. good nutritional status

- Means poor outcome indicator values - e.g. poor nutritional status

It may even be possible to derive some conclusions where there is no difference in delivery, but where differences in socio-economic status still exist. The possibilities are given in Table I.7b. Here example 4 indicates there is an inadequate effect of the program and it should be further examined. Example 5 indicates that the program may be having an adequate effect, although it is possible that socio-economic status does not account for any differences. Example 6 indicates that the program is having no effect and should be further examined or discontinued. Example 7 is in practice unlikely to occur. Such tabulations give useful insights into the program's adequacy, and also raise questions on targetting and delivery, as discussed in the next section.

Definitions of population groups involved.

Both for planning and evaluation, it is important to distinguish between different population groups. The main groups of concern are as follows:

(a) The total population in the program area;

(b) The population targetted by the program;

(c) The population in need of better nutrition - called "needy" here;

(d) The population receiving benefits from the program - called "recipients" here.

In some cases additional sub-groups might need to be considered, e.g.:

(e) The needy population who could benefit from the program; examples of needy who could not might be malnourished children in a supplementary feeding program whose nutritional problem is due to malabsorption and not to inadequate food intake.

If program staff have contact with the recipients, obtaining data on these may be relatively easy. This would be the case in the example of a feeding program, but maybe not in, say, a water supply project. If outcome data are available from recipients these can, to a limited extent, substitute for survey data on the population as a whole.

The distinction between population groups allows construction of a series of 2 x 2 tables which lead to some important indicators for planning and evaluating targetting, as shown in Table I.8. In planning, the two important indicators are:

TABLE I.8
Quantifying Target Groups, "Needy", and Program Recipients 1/

A.

		Needy? (e.g. malnourished) YES	NO
Targetted?	YES	a_1	b_1
	NO	c_1	d_1

Planning (Pre-Program)

Base-line Data

"Planned Focusing" [1]: Proportion of total targetted that are needy = $\frac{a_1}{a_1 + b_1}$

"Planned Coverage" [2]: Proportion of total needy that are targetted = $\frac{a_1}{a_1 + c_1}$

B.

		Targetted? YES	NO
Recipient?	YES	a_2	b_2
	NO	c_2	d_2

- Program data

Evaluation (During Program)

Survey data (directly)

"Leakage"
- Proportion of total recipients who are targetted = $\frac{a_2}{a_2 + b_2}$
- Proportion of total recipients not targetted = $\frac{b_2}{a_2 + b_2}$

"Delivery"
- Proportion of total targetted not recipients = $\frac{c_2}{a_2 + c_2}$
- Proportion of total targetted who are recipients = $\frac{a_2}{a_2 + c_2}$

C.

		Needy? YES	NO
Recipient?	YES	a_3	b_3
	NO	c_3	d_3

- Program data

Evaluation (During Program)

Survey or non-program data (directly)

"Actual Focusing" Proportion of total recipients who are needy = $\frac{a_3}{a_3 + b_3}$

"Actual Coverage" Proportion of total needy who are recipients = $\frac{a_3}{a_3 + c_3}$

(With delivery exactly as targetted, recipients = targetted, and this table is exactly as Table A)

1/ In the epidemiological terms, [1] focusing is equivalent to positive predictive value; [2] coverage is equivalent to sensitivity.

(a) Proportion of total targetted that are needy (needy targetted/ total targetted) which indicates the degree of "planned focusing" of the program towards nutrition.

(b) Proportion of total needy that are targetted (needy targetted/total needy), which reflects the **"planned coverage"** of the program.

The concepts of coverage and focusing have commonsense meanings, both for planning and evaluation. **Coverage** is a basic value that needs to be manipulated for different program designs (equivalent to sensitivity in the epidemiological literature, see Habicht, Meyers and Brownie 1982): evidentally the aim is to optimize coverage. Costs should be put into this calculation. **Focusing** (equivalent to positive predictive value, see Habicht, Meyers and Brownie 1982) is a less familiar concept; at a minimum, if targetting is to focus resources, focusing should be greater than the prevalence in the population as a whole (i.e. proportion of needy in the targetted population should be greater than proportion of needy in the population as a whole; the same could apply - but is seldom to our knowledge done - for any evaluation of "poverty-orientation"). Procedures for choosing appropriate indicators and their screening levels to identify proportion of needy, and for efficiently deciding on cut-off points to define needy, are discussed in Habicht, Meyers and Brownie 1982.

For evaluation, the delivery is compared with the targetting and with degree of need, to generate further indicators as shown in Table I.8. This requires determining whether the recipients were in fact targetted, and if they are needy (e.g. malnourished).

An intermediate stage comparing targetted with recipients, gives indicators of **delivery** (Table I.8b) e.g.

(c) Proportion of total targetted who are recipients, which should be 100% if the program is fully implemented.

and of **leakage**, e.g. as

(d) Proportion of total recipients who are targetted, or conversely proportion of total recipients who are <u>not</u> targetted. These should be 100% and 0% respectively if there is no leakage to non-targetted groups.

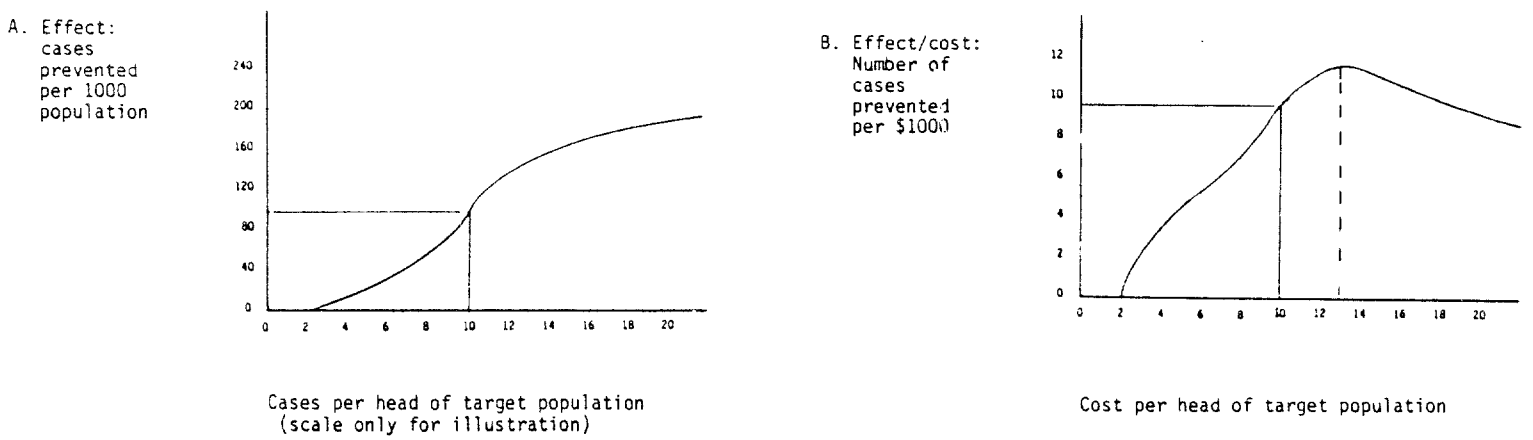
If there is full implementation and no leakage, then the **"actual focusing"** and **"actual coverage"** are the same as those planned - see Table I.8c. If there is deviation from the plan, then one way of assessing this is to calculate these "actual" indicators, comparing "needy" with "recipients". Again, for example, at a minimum actual focusing (recipients needy/all recipients) should be greater

than the population prevalence of needy. Even without knowledge of costs, such indicators could give useful means of evaluating process; with costs as discussed in the next section, they could lead to decisions as to whether the program is within the range likely to give an adequate or acceptable outcome, even if the expected effects on recipients were achieved. A worked example is given in Mason et al. (1982 Chapter IV).

If data are available on needy, recipients and targetted populations from baseline studies, then some conclusions can also be drawn on outcome during program implementation based only on outcome data on the recipients. This is so if the assumption can be made that the change in outcome variables is likely to be small compared with that in recipients, and if baseline (pre-program) data are available. When this applies, it reduces the need for population surveys for evaluation. This theory is also given in Mason et al. (1982 Chapter IV).

FIGURE I.1

EFFECT/COST CURVES



Effect/cost

Cost-benefit and cost-effectiveness analyses are commonly used for assessing many type of programs, both during planning and for evaluation. In the present case, since a monetary figure cannot reasonably be put on outcome, cost-effectiveness is the more suitable approach. It is however not often used in evaluating nutrition or feeding programs, and a major advance in these evaluations could be made by much more systematic introduction of the techniques and thinking involved. These do not necessarily depend on accurate data, and indeed some form of cost-effectiveness thinking is implicit in the planning of almost any program: that there is a level of expenditure per unit of expected outcome that would not be worth it is almost always in the back of someone's mind. We consider that the summary parameter of effect per unit costs (which goes to zero when there is not effect) is a useful start, and this is the one mainly discussed here.

A dose-response type of curve relating effects to cost is likely to apply to intervention programs. This is familiar in economics (as in total product and utility curves, etc.), but no often considered for nutrition programs. This means that the relationships shown in Figure I.1 are likely to apply. Probably, there is as yet insufficient data to put a scale on the X-axis, but some research on existing data might allow hypotheses to be put forward. In this hypothetical example, a cost per head of the target population of around $13 gives the maximum cost-effectiveness calculated as number of cases prevented per thousand dollars (Fig. I.1B); but this rate of expenditure gives less than the maximum overall effect Fig. I.1A). The two curves are directly related: for example at $10 per head expenditure, if 100 cases per thousand population are prevented (A), this is 100 cases per $10,000 or 10 cases per thousand dollars (B). The effect/cost in (B) for any value of cost per head is equal to the total effect as can be read off in (A), divided by the corresponding cost per head. Put another way, the height of the curve in (B) at any value of cost per head is the slope of the line joining the origin to the corresponding point on the curve in (A).

One important advantage of such methods would be to allow assessment as to whether the level of effort in a program is at least in the range in which an outcome effect could be expected, taking account also of the level of malnutrition in the target group. It is our impression that often little effect of a program could in reality be expected because the level of expenditure relative to

the expected dose-response is not even in the right range - being far too low. Something like this idea has been referred to as "situation assessment" (Riecken 1979, page 368).

Another application of calculating effects per unit cost is to define the extent to which an accurate assessment of outcome is needed. For example, (using relationships similar to those in Figure I.1) it might be postulated that a change from 20% prevalence to 10% prevalence after the treatment is the maximum feasible (e.g. from 200 malnourished in a population of 1,000 to 100 malnourished) at a cost of say $10 per head (i.e. $10,000 for the population of 1,000). This is equivalent to proposing an effect per unit cost of 10 cases prevented or rehabilitated per $1,000. Clearly, this should have been regarded as good value for money at the stage of planning the project. Similarly, no change would mean that effect per cost was zero. Somewhere between these two a level of change could be set below which it was regarded that the program's resources were not being well spent, for various reasons which could relate to targetting, type of activity, adequacy of delivery, etc. For example, rehabilitation of 5 cases per $1,000 could be regarded as the minimum effect/cost ratio acceptable. This means that the maximum acceptable post-program prevalence is 15% (i.e. a maximum of 150 malnourished in the population of 1,000). In this case, it is only necessary to know whether the with-program prevalence is above or below the adequacy cut-off point of 15%.

Appropriate Indicators for Different Objectives

So far we have not defined or commented on specific potential outcome indicators, and have used nutritional status as measured by anthropometry as the general example. This was in line with our brief for contributing to the MIT workshop, and with our view that the major problems lie in designs of evaluation rather than in the measurements to be taken. However, it is clear that some summary of appropriate indicators at this stage would be useful; moreover this reinforces the point that expected outcome needs to be related to program design from very early on in either planning or evaluation. At the MIT meeting, we made two suggestions for summarizing suitable indicators. The first was to put together what was known quantitatively about the use of different indicators, in particular, their **responsiveness** to certain interventions. When it became clear that these objective criteria for choosing outcome indicators were only very sketchily known, we adopted a second approach which was essentially to ask the participants for their expert opinion as to suitability of indicators, by

program type and objectives. These results have been compiled, with the addition of our own opinions. The responsiveness results are set out in Table I.9, and the more subjective opinions in Table I.10. Although Table I.9 is unsatisfactory in the sense that only a few indicators have been objectively evaluated, it also serves to demonstrate the sort of evaluation of indicators that now needs to be undertaken much more widely to establish a firm basis for selection of these in the future. In the meanwhile - and not as a substitute - the guidelines in Table I.10 may be of some help.

The relation between indicators and objectives often needs to be clarified. Sometimes the indicator precisely measures the objective. A feeding program aims to increase the weight gain in a target population of pre-school children, and this weight gain itself is measured. In this case, the responsiveness of the indicator is equivalent to the effectiveness of the program. In other circumstances the indicator is a proxy for the main objective: a feeding program aims to increase the food intake of a target group of children, but the food intake itself is not open to measurement, so anthropometry is used as a proxy for the food intake. In this case we need an indicator that responds to increased food intake; thus for example Habicht and Butz (1979) showed that height gain is more responsive than weight gain (in the statistical sense of greater significance) and thus a better proxy for food intake. Such relations between the indicator and the objective need to be established in advance.

As shown in Table I.9, some relevant data were obtained to allow comparison of the responsiveness of different indicators. As might be expected, the responsiveness is very much higher in a clinical setting than in a field trial. This is presumably because of the greater severity of malnutrition being treated and to a lesser extent due to increased precision and reliability of measurement. The field trials are likely to be more important for the present purposes, both because the programs to be evaluated more closely correspond to these, and also because the required measurements shown to be responsive in a clinical setting mostly require laboratory facilities - in this case, for example for immunology, quite extensive facilities. (For comparison, a hypothetical improvement of 10 percentage points in mean weight/age for a treatment over a control group, e.g. from 75% mean weight/age to 85% mean with the common standard deviation of 10 percentage points, means a responsiveness of this indicator in this situation of 0.5.) The results from Martorell et al. (1980) support the use of weight and height in pre-school children over circumferences or skin folds; this is similar to

TABLE I.9

Mean Indicator Response to Supplementary Feeding

Field Trials

Type of Malnutrition	Type of Analysis	Indicator	Age	Duration of Suppl.	% Suppl. Diet	Deficit rel. to std.	Response to Suppl.	Pooled SD	Responsiveness = ½ (Respon/SD)^2
PEM	Suppl. vs. Control	**Attained:** Wt.	36 mo.	36 mo.	17% - Cal. 36% - Pro.	4.5 kg. (Denver)	0.9 kg.	1.3 kg.	0.24
		Ht.	"	"	"		2.3 cm.	3.9 cm.	0.17
		Arm Circum.	"	"	"		0.35 cm.	0.9 cm.	0.08
		Triceps Skinfold	"	"	"		0.15 mm.	1.1 mm.	0.01
		Subscapular Skinfold	"	"	"		0	1.1 mm.	0
		Ref: Martorell, et al. 1980. Nutr. Rep. Int. 21: 219-230							
Vit. A	Pre & post intervention	Serum Retinol	Pre-School	1-2 yr.	>100% Vit. A req.	Std. ≥ 20 mcg/dl	12.3% decline in prevalence values < 20 mcg/dl	34.7%	0.06
		Ref: Arroyave, et al. 1979 Sci. Pub. 384 PAHO							

Type of Malnutrition	Type of Analysis	Indicator	Age	Duration of Suppl.	% Suppl. Diet	Deficit rel. to std.	Response to Suppl.	Pooled SD	Responsiveness = ½ (Respon/SD)^2
Iron deficiency anemia	Intervention vs Control Group	Hgb (g/dl)	9 mo.	6 mo.	15 mg. Fe + 100 mg. Ascorbic Acid per 100g. full fat milk powder	1.21	1.07 g	1.0 g	0.57
		Sat (%)	"	"		8.2	4.8%	6.0%	0.32
		FEP (mcg/dl RBC)	"	"		39	26 mcg	33 mcg	0.31
		% children with Hgb< 11.0g/dl	"	"			27.[illegible]%	38.4%	0.25
		HgB	15 mos.	9 mo.	(as above)	0.92	1.02 g	.94 g	0.59
		Sat	"	"		6.7	7.2%	8.0%	0.405
		FEP	"	"		38	24 mcg	41 mcg	0.17
		% children with Hgb< 11.0g/dl.	"	"			26.2%	32.8%	0.32

Ref: E. Rios, et al. forthcoming. Prevention of iron deficiency in infants by milk fortification. In Nutrition Intervention strategies, B. Underwood ed.
A. Stekel: personal communication 1981.

Clinical Trials

Type of Malnutrition	Type of Analysis	Indicator	Age	Duration of Suppl.	% Suppl. Diet	Deficit rel. to std.	Response to Suppl.	Pooled SD	Responsiveness = ½ (Respon/SD)^2
PEM	Response to protein supplement	VO_2 max. /kg. (direct measure)	39 ± 9.7 yrs.	2½ mo. (80d)	Protein from 5.6% - 19.6% of calories	20	9.7	5.14	1.89
	difference between means	VO_2 max. L/min. (direct measure)	39 ± 9.7 yrs.	"	"	1.49*	0.75	0.314	2.85
		heart rate response to a workload of 250 kgM/min.	39 ± 9.7 yrs.	124d	2240 kcal./d + 357 kcal 100 Gm protein	55*	30	6.63	10.24

* with respect to valve in general from workers of the same region (normals)

Ref: 1. Barac-Nieto et al. 1980. Am. J. Cl. Nutr. 33:2268-2275
2. Maksud et al. 1976. Eur. J. Appl. Physiol. 35:173-182.
3. Spurr et al. 1979. Am. J. Clin. Nutr. 32: 767-778

PEM	Response to R_x clinical grp. only	Serum prealbumin mg/100 ml	18-30	22 days	% not given; Nido + Nesmida in↑ amt.	All with clinical PEM	PA-16mg	2.79	16.4
		Serum albumin g/100 ml					Alb-1.51g	0.36	8.80

Ref: Ingenbleek et al. 1972. Lancet ii:106

Type of Malnutrition	Type of Analysis	Indicator	Age	Duration of Suppl.	% Suppl. Diet	Deficit rel. to std.	Response to Suppl.	Pooled SD	Responsiveness = ½ (Respon/SD)^2
PEM	Normal vs. pre & post PEM	serum alb. TBA RBP plasma retinol	18-30 months	22 days	↑ to plateau of 3.5 g prot. & 150 kcal/kg BW/d	52.8% of control 28.5% " 31.9% " 27.4%	1.48 g/100 ml 15.93 mg/100 ml 3.79 mg/100 ml 30.64 g/100 ml	0.38 2.79 0.80 7.49	0.16 16.3 11.22 8.37
Ref: Ingenbleek et al. 1975. Clin. Chim. Acta 63:61									
Kwashiorkor	Comparison of pre & post intervention	3 rd. component of complement C_3 (mg/100 ml.)	6 m/o - 6 y/o	2 wks.	from: 0.8g prot. 88 kcal/kg/d to: 3.5-4g. prot. and 140 kcal/kg/d plus multi-vitamins	34 mg.	22 mg.	4 mg.	15.13
Ref: Neumann et al. 1975. Am. J. Clin. Nutr. 28:89-104									
PEM	Comparison of pre & post intervention	% T - lymphocytes in blood	Children	6-16 wks.	"correction of deficit"		37%	9.2%	8.1
Ref: Chandra 1974. Brit. Med. J. 3: 608-609									

Type of Malnutrition	Type of Analysis	Indicator	Age	Duration of Suppl.	% Suppl. Diet	Deficit rel. to std.	Response to Suppl.	Pooled SD	Responsiveness = ½ (Respon/SD)^2
PEM	Comparison of pre & post inter-vention	% T lymph-ocytes in blood	1-5 y/o	29 days	1g prot. 100 kcal/kg/d 4g prot. 175 kcal/kg/d		33.0%	17.0%	1.9
				Ref: Kulapongs 1977 et al. in Malnutrition and the Immune Response, R.M. Suskind ed. New York: Raven Press, 99-103					
Iron Deficiency Anemia	Comparison of pre & post inter-vention	Bactericidal capacity of PMN's leucocytes	1-8 y/o	1 dose parenterally	Iron		.33	.14	2.8
				Ref: Chandra 1973. Arch. Dis. Child. 48:864-866					

TABLE I.10

Outcome Indicators considered useful for evaluation of different programs and their objectives

PROGRAM TYPE OBJECTIVES	INDICATORS[1]	COMMENTS
1. **Pre-school child programs**		
(a) Reduce PEM	A. Wt. and ht. changes Ht/age, wt/age, wt/ht	Longitudinal measurement of weight and height preferable if can be organized. Height shows best response in long term. Wt/age, etc. used when population is assessed at various points in time. Ht/age, wt/age most useful. Wt/ht not sensitive unless substantial deficit in population to begin with. Constraints are equipment availability and accuracy, trained personnel. Ht measurements may be more difficult than wt. Growth charts recording wt/age are often available.
	B. Clinical signs Dietary intake Limb circumferences and skinfolds	Prevalence of clinical signs such as oedema is usually too low to be useful except in emergency situations. Others, such as hair changes, are less specific and change slowly. Subjective recording of degree of wasting unreliable. Dietary intake measurements on individuals are very expensive, and require much time, training, and supervision; not widely used with success. Circumferences may be useful for screening, but are probably less responsive to change. Skinfolds require more skill in measurement, and have no additional advantage.
(b) Reduce morbidity (with e.g. MCH)	A/B Prevalence of diseases No. of episodes Duration longly	No evidence that morbidity responds to food intake alone, but relevant in context of e.g. MCH services. Problems longly in collection and interpretation. Episodes defined by symptons not disease: e.g. diarrhea, cough, etc. Immunization scars can be used when appropriate.

---

[1] A. Widely used and recommended for evaluation
B. Less often used - mainly for research studies

PROGRAM TYPE OBJECTIVES	INDICATORS	COMMENTS
(c) Reduce infant and child mortality	B. Infant and child mortality rates	Changes in child mortality have been shown in pilot studies to be responsive. However, all of these demonstrated changes required detailed records of specially established vital registration and extensive supervision and continuing research inputs. A general opinion is that mortality rates are difficult to measure accurately, expensive, and hence not generally suitable for routine evaluation.
2. School feeding programs		
(a) Improved nutritional status	A. Heights & weights longitudinally B. Other anthropometric, biochemical tests	Comments much as for pre-school, longitudinal measurements easier here. Other tests seldom worthwhile; as for pre-school.
(b) Improved school enrollment and attendance	Records of enrollment and attendance.	No results presented, but presumably problems are mainly organizational.
(c) Improved school performance	School performance tests	See relevant chapter.
(d) Income transfer; increased food intake	B. Consumption, income expenditure	Budget and consumption surveys needed; possibly obtainable from household budget surveys if program participation data obtained on questionnaire; otherwise research rather than routine application.
3. Food for work programs		
(a) Increase productivity	B. Physical activity, energy expenditure,	These can be measured in research. Not much experience.
(b) Increase income and/or food consumption.	B. Household surveys	Suitable for research, not routine. Not much experience.
4. Relief in emergencies		
(a) Rehabilitation: children	A. Anthropometry Clinical signs	See pre-schoolers; weight gain and wt/ht are useful; clinical signs (e.g. oedema, signs of severe wasting) also useful.
(b) Rehabilitation: adults	A. Weight gain	Could be used; not reported however.

PROGRAM TYPE OBJECTIVES	INDICATORS	COMMENTS
5. **Supplementary feeding for mothers**		
Reduce: risks at birth, low birth weight, subsequent infant mortality	A. Weight gain during pregnancy Birth weight	These require extensive health service coverage and good record keeping; given this, "A" indicators, particulary birth weight, are fairly straightforward.
	B. Perinatal and infant or mortality rates	IMR's require additional follow-up, vital registration linked to program delivery data - see 1 (c).
6. **Vitamin A fortification and distribution programs**		
Reduce vitamin A deficiency	A. Clinical signs (eyes) B. Serum retinol	Clinical signs may be complicated by other causes of eye changes. Decrease in low end of serum retinol distribution is internationally accepted procedure (IVACG).
7. **Iron fortification and distribution programs**		
Reduce iron deficiency	B. Clinical signs Hemoglobin Serum ferritin Transferrin saturation	Anemia can be estimated from e.g. color of optical mucosa, but has subjective error. Biochemical tests require blood sample; hemoglobin not specific for iron deficiency anemia; transferring saturation more demanding but more specific.
8. **Iodine fortification and injection programs**		
Reduce goiter; cretinism. Increase mental capacity in iodine deficient areas.	A. Goiter prevalence, by size B. Cretinism and IQ tests	Goiter can be estimated by trained health personnel. IQ tests can be applied to children after mothers supplementation.

the conclusions of Habicht and Butz (1979). For vitamin A, serum retinol measurements are shown to be reasonably responsive (e.g. equivalent to weight) from the data of Arroyave et al. (1979). In field trials of iron supplementation hemoglobin concentration (but not hematocrit) and transferrin saturation came out as responsive at least in the two sets of results made available. Such responsiveness characteristics provide guidance as to required sample size (see next section), measurement accuracy and reliability, etc., in designing evaluations. Where possible, these characteristics of the further indicators advocated should be given in the appropriate chapters subsequently in this book. There is an urgent need for research to establish the responsiveness characteristics of other indicators in a manner equivalent to that given in Table I.9.

Table I.10 reflects the state-of-the-art of evaluation of eight types of nutrition programs in terms of outcome indicators. The most widely used and trusted measurements are clearly anthropometric, where these correspond to program objectives, either directly (e.g. 1(a)) or as proxies (e.g. 5, where increased birthweight is considered to decrease subsequent mortality risk, even if the subsequent mortality cannot be measured). The general theme of the conclusions shown in this table are that most measurements are difficult to obtain with any reliability in routine programs - this applies to morbidity and mortality data, biochemisty, and food consumption estimates. These measurements are therefore more appropriate within experimental studies, as discussed earlier in this chapter. This extensive reliance on anthropometry in practice for routine operational program evaluations was not obvious until this table was drawn up. Indeed it may not be desirable, and emphasizes the need for further assessment of alternative outcome indicators. But it does seem to represent the consensus of those at the workshop. The absence of an extensive list of well-tried outcome indicators for routine evaluations also points up the value of process evaluation, and the need to focus on only those programs where an effect on outcome is likely to be detectable; this is the theme of much of the next chapter.

Note on Sample Size: (incorporated from a note prepared by M. Guzman)

The question of sample size merits careful consideration in attempts to evaluate the results of any administered treatment or treatments.

In the first place, the investigators must define carefully the unit of reference for which the sample size is to be estimated, clearly differentiating "observational units" from the "unit of interest" for the evaluation. It is the latter which is the principle determinant of sample size. It is made up of a cluster of observational units. In other words, although information may be collected from individuals (observational units), the evaluation of effects may focus and center interest in aggregates of individuals who constitute say, families - a cluster of observational units.

Whatever the "unit of interest", the number of such units (sample size) should be estimated under pre-specified conditions of accepted risk a) of detecting an effect when in fact it does not exist and b) of not detecting the effect when it does exist. In the procedures for the statistical testing of specific hypotheses relating to treatment effects, the relative frequency (probability) of occurrence of the first kind of error is used to define the level of significance for performing the test, while the frequency of non-occurrence of the second kind of error is used to define the power of the test (frequency of correct detection of effects).

Under these premises, and provided the investigator can provide a priori information on the magnitude of the minimum treatment effect (expected result of a control-treatment difference) he considers worth identifying with concomitant information that relates to the variability (standard deviation) of the response under consideration, it is possible to estimate the approximate size of the sample required under this particular set of conditions to detect the treatment effect (see Cohen 1969 for textbook treatment of this issue).

The procedure for estimating sample size for comparison of independent samples utilizes the relation:

$$n = \frac{K^2\, 2s^2}{d^2} = \frac{K^2}{\frac{1}{2}\,(d/s)^2} = \frac{K^2}{f^2 \times \text{Responsiveness}}$$

- where s is an estimate of the standard deviation of the variable under consideration (see Table I.9)
- d is the difference to be detected; this should be a fraction (f) of the responses shown in Table I.9, below which one is indifferent to whether there is a response.
- K^2 is a multiplier value for different levels of significance and various associated powers in testing as follows:

VALUES OF K^2

Power (0/)	Level of Significance Two - tailed test 1%	5%	10%	Level of Significance One - tailed test 1%	5%	10%
80	11.7	7.9	6.2	10.0	6.2	4.5
90	14.9	10.5	8.6	13.0	8.6	6.6
95	17.8	13.0	10.8	15.8	10.8	8.6

Source: Snedecor and Cochran, 1980.

In closing it should be restated that the above procedure is only an approximation and is usually an underestimate of the required sample size. Its indiscriminate application may lead to absurd answers. It is advisable, therefore, that the question of sample size always be considered in the context of each particular situation and with proper statistical consultation. Ultimately, the successful estimation of sample size is the result of experience which bridges across the realms of art and science.

Part I - References

Arroyave, G.; Aguilar, J.R.; Flores, M.; and Guzman, M.A. 1979. Evaluation of Sugar Fortification with Vitamin A at the National Level. Scientific Publication 384. Washington, D.C.: PAHO.

Beaton, G.H. and Ghassemi, H. 1979. Supplementary Feeding Programmes for young children in developing countries. Mimeographed. Report prepared for UNICEF and the ACC sub-committee on Nutrition of the United Nations.

Beghin, I./FAO 1980. Selection of specific nutritional components for agricultural and rural development projects. Mimeographed. Antwerp, Belgium: Nutrition Unit, Institute of Tropical Medicine.

Casley, D. and Lury, D. 1982. A Handbook on Monitoring and Evaluation of Agricultural and Rural Development Projects. Baltimore: Johns Hopkins Press.

Chernichovsky, D. 1979. The economic theory of the household and impact measurement of nutrition and related health programs. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et al., pp. 227-67. New York: Plenum Press.

Cohen, J. 1969. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press.

Cook, T.D. and Campbell, D.T. 1979. Quasi-Experimentation. Boston: Houghton Mifflin.

Davis, C.E. 1976. The effect of regression to the mean in epidemiologic and clinical studies. Am. J. Epidemiol. 104: 493-98.

Deboeck, G. and Kinsey, B. 1980. Managing Information for Rural Development: Lessons from Eastern Africa. World Bank Staff Working Paper No. 379. Washington, D.C.: World Bank.

Drake, W.D.; Miller, R.I.; and Humphrey, M. 1980. Final Report: Analysis of Community-Level Nutrition Programs. Project on Analysis of Community Level Nutrition Programs. Vol. I. Washington, D.C: USAID-Office of Nutrition.

Furby, L. 1973. Interpreting regression toward the mean in development research. Developmental Psychology 8: 172-9.

Gwatkin, D.R.; Wilcox, J.R.; and Wray, J.D. 1980. Can health and nutrition interventions make a difference? Overseas Development Council Monograph No. 13. Washington, D.C.: ODC.

Habicht, J-P. 1979. Assurance of quality of the provision of primary medical care by non-professionals. Social Science and Medicine 13B (1): 67-75.

Habicht, J-P. and Butz, W.P. 1979. Measurement of health and nutrition effects of large-scale nutrition intervention projects. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et. al., pp. 133-82. New York: Plenum Press.

Habicht, J-P.; Meyers, L.D.: and Brownie, C. 1982. Indicators for identifying and counting the improperly nourished. American Journal of Clinical Nutrition, forthcoming.

Judd, C.M. and Kenny, D.A. 1981. Estimating the effect of social interventions Cambridge: Cambridge University Press.

Kielmann, A.A.; Ajello, C.A.; and Kielmann, N.S. 1980. Evaluation of Nutrition Intervention Projects. Final Report to the Documentation Coordinator, TA/PPU/EUI, Technical Assistance, AID. Mimeo. Washington, D.C.: USAID.

Klein, R.E.; Read, M.S.; Reicken, H.W.; Brown, Jr., J.A.; Pradilla, A.; and Daza, C.H. 1979. Evaluating the impact of nutrition and health programs. New York: Plenum Press.

Martorell, R., Klein, R.E.; and Delgado, H. 1980. Improved nutrition and its effects on anthropometric indicators of nutritional status. Nutrition Reports International 21: 219-230.

Mason, J.B.; Habicht, J-P.; Tabatabai, H.; and Valverde, V. 1982. Nutritional Surveillance. Forthcoming - presently mimeo. Ithaca, NY: Cornell University.

Poister, T.H. 1978. Public Program Analysis. Applied Research Methods. Baltimore: University Park Press.

Riecken, H.W. 1979. Practice and Problems of Evaluation: A Conference Synthesis. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et al., pp. 363-86. New York: Plenum Press.

Sahn, D.E. and Pestronk, R.M. 1981. A Review of Issues in Nutrition Program Evaluation. AID Program Evaluation Discussion Paper No. 10. Washington, D.C.: USAID, Offices of Nutrition and Evaluation.

Saretsky, H. 1972. The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan 53: 579-81.

Snedecor, G.W. and Cochran, W.G. 1980. Statistical Methods. 7th ed. Ames, Iowa: Iowa State University Press.

Wray, J.D. 1970. Malnutrition is a problem of ecology. Bibl. Nutr. Diet. No. 14: 142-60.

Wray, J.D. 1973. Twenty questions: a checklist for planning and evaluating nutrition programmes for young children. In Nutrition Programmes for Preschool Children, D.B. Jelliffe & E.F.P. Jelliffe, eds. pp. 323-333. Zagreb: Inst. Public Health.

PART II
STAGES OF EVALUATION OF ON-GOING PROGRAMS

Introduction

The principles for the evaluation of on-going programs were set in Part I. It was argued there that proper evaluation of such programs should de-emphasize the researcher's concern with causality and prediction and seek to provide answers to questions likely to be of interest to managers, administrators and funders. It was also pointed out that evaluation should proceed in a sequence of stages so that easier and cheaper answers are found first before embarking on more elaborate types of investigation which may prove to be either unnecessary or impossible. This chapter takes up this sequence of stages in the evaluation of on-going programs.

Since finding answers to all the possible questions from evaluation - for example as given in Table I.4 - is an extensive and costly task, it is necessary to delineate what is needed and what can be done with given resources. These refer to the needs of managers, administrators or funders, and researchers, who require different degrees of certainty or plausibility for their answers and conclusions, on the one hand; and to the feasibility and costs of carrying out various types of investigation to respond to these needs on the other.

In practice, therefore, the outcome and its possible association with program delivery should be examined during the later stages of the evaluation process. Before getting this far a series of stages should be followed to determine whether it is necessary to examine outcome at all; and if so what associations between program delivery and effect should be looked for. The rest of this chapter, therefore, sets out a sequence of stages that starts with the less expensive investigations and only proceeds to examine whether cause-effect relations can be shown after studies on the program design and its implementation have shown these effects to be likely to exist. These stages are as follows:

Stage 1: Preliminary Tasks: to answer the questions of evaluation for whom, of what, and in what context

Stage 2: Evaluating the Plan of the Program

Stage 3: Evaluating the Implementation of the Program

Stage 4: Evaluating the Gross Outcome of the Program

TABLE II.1

STAGE 1: PRELIMINARY TASKS

1. **Decide:**

   - Who is doing the evaluation?
     (e.g. outsiders, management, funders)

   - For whom?
     (management, administrators and funders, research bodies)

   - For what purpose?
     (see Table I.2) e.g.:

     - continuation or modification of delivery of the program?

     - replicate the program?

     - estimate the net effects of the program?

2. **Reach consensus on the objectives of the program.**

   - Make underlying objectives explicit and resolve contradictions.

3. **Scouting**

   - Get initial impression

   - Make explicit how the program is expected to achieve its effects

4. **Plan the evaluation**

Stage 5: Evaluating the Net Outcome of the Program

Stage 6: Presentation and Dissemination of Evaluation Results

Stage 7: Move to built-in evaluation

A summary of each of these stages is given in each section as a table: these can be reviewed together as a summary of the procedure set out in this chapter.

STAGE 1: PRELIMINARY TASKS

(See Table II.1)

A first question should be: who is doing the evaluation, for whom, and for what purpose. Those responsible for carrying out the evaluation could be outsiders, management, the funding body, or combinations of these. The users of the evaluation results could be management, administrators and funders, research bodies (see Table I.2), or all of these; these should determine the purposes of the evaluation. If the purpose is to provide information to make decisions, the actual options to be decided upon should be specified. The purposes in turn should dictate the questions and issues addressed, and hence the decisions to be taken, again as outlined in the previous chapter. These purposes in effect constitute terms of reference and should be made explicit at the outset. At the same time the degree of certainty needed to make the consequent decisions, and hence the methods to be used, must be decided.

The next initial step is to agree upon the program objectives because an evaluation should result in judgements about how well a program is meeting its objectives. Those involved in programs often have very different perceptions of program objectives, and these perceptions may change markedly over time. Therefore, discussions must examine program objectives and resolve contradictions that may exist among these, so that a consensus on the objectives of the program can be set out explicitly.

In practice, it is essential then to get a feel for the program. This has been referred to as scouting (Riecken 1979, p. 369), meaning essentially going and talking to people in the program area, both to those responsible for the program, and to those participating in or benefitting from it. This is worth several days' or even weeks' effort, providing initial impressions as well as a basis for proceeding. It should also involve drawing up some sort of conceptual model to explain how the program is supposed to achieve its effects - if this is not already available in the program documents - which means making the processes of the project explicit. Such a model also helps to identify the constraints to implementation, and the possible variables confounding apparent effects. An

TABLE II.2

STAGE 2: EVALUATING THE PLAN OF THE PROGRAM

1. **Examine overall objectives**

   - What are they?
   - Are objectives specified in terms of:
     - quantities of inputs?
     - target groups (numbers, characteristics)?
     - permissible deviations in targetting and delivery?
   - Are inputs compatible with expected outcomes (was there a sound basis for the program design)?
   - If above cannot be specified - **stop**

2. **Evaluate implementation objectives**

   - Budget, work-plan
   - Planned expenditure per head
   - Compare with expected effect per unit cost

3. **Evaluate targetting objectives**

   - Who, how many?
   - Calculate planned focusing and coverage (see Table II.8)

4. **Evaluate outcome objectives**

   - Apply to who (recipients, targetted, total population)?
   - Do objectives match inputs?
   - Is planned effect/unit cost reasonable?
   - Define adequate effect.

5. **If adequate effect not expected from evaluation of plan**

   - **stop**

example of such a model based on the sequence of events in a food aid program could be as follows: food is delivered to a central warehouse; it is distributed to MCH clinics in target areas; the food is given monthly in specified quantities to mothers of children of less than 80% weight for age; at the same time malnourished children come to the MCH clinic and are identified; the mothers give their children the supplementary food, which adds to the previously existing intake, rather than replaces it, to a specified extent; this in turn leads to improved nutritional status of the target group.

From such information the evaluation itself should be planned, specifying the different stages (e.g. along the lines suggested here.) Constructing a flow chart of the sequence of steps involved in each stage may be useful. An important decision even at this stage is whether it will be feasible to collect fresh data, either by case-study or sample survey, should the next stages indicate that this is worthwhile. This data collection is expensive, and resources should be concentrated on those programs where this is necessary. Stages 2 and 3 (usually) do not require surveys and efforts devoted to these earlier stages may pay off by making further data collection unnecessary. Stages 4 and 5 usually will require fresh data collection. However, whether these stages will actually be reached depends on the outcome of preceeding stages, and cannot yet be decided.

STAGE 2: EVALUATING THE PLAN OF THE PROGRAM

(See Table II.2)

The evaluation of the plan of the program seeks to answer the question: what effect was expected, or could reasonably have been expected, if the program were implemented as planned? A systematic approach to this question involves a number of steps.

The overall objectives of the program must be explicitly stated in Stage 1. The evaluation of these objectives then is concerned with checking their feasibility given the resources available and their intended effect per unit cost. The initial steps concern the following questions:

- Are objectives specified in terms of:
  - quantities of inputs
  - target groups (numbers, characteristics)
  - permissible leakage (percentage non-targetted recipients) and delivery (proportion targetted who are recipients)

- Are planned inputs generally compatible with expected outcomes?
- If the answer to the first two is negative, can the objectives now be specified in terms of target groups, leakage, etc., and expected outcomes? If not - **stop.**

The next step is to systematically evaluate **implementation** (input) objectives, **targetting** objectives (including permissible delivery and leakage), and **outcome** objectives, including determination of **adequate** levels of achievement of these.

Evaluating Implementation Objectives

This step should first check that sufficient details on supplies, services, costs, etc. as programmed are available to provide a basis for evaluating the overall extent to which the program has been implemented. The program budget and work plan should provide the needed information. If not, further inquiries may be needed.

At this point, calculate the planned expenditure per head of the target population. Due allowance must be made for both direct and visible costs - e.g. food, transport, additional administration - and less visible costs - e.g. the opportunity cost of the time of personnel who would be occupied otherwise in the absence of the program. In certain projects the planned level of expenditure per head may be too low to permit a realistic expectation of any detectable (or important) effect of the program on nutrition. As noted in Part I, there is a scarcity of empirical data on actual levels of expenditure, certainly in relation to effect. However, it is important to check whether there is any basis in experience under the prevailing conditions to suppose that the levels of expenditure planned can have the expected effect. In the absence of local data, the figures quoted by Gwatkin et al. (1980) and Beghin (1980) may provide some guidance. As a very rough rule of thumb, it seems to us that an expenditure below a minimum of say $1 - $10 per recipient is very unlikely to produce any measurable effect on outcome. Once such estimates are made for the program, they need to be:

- checked with the agencies concerned;
- put together with the evidence for expected effects, to reach approximate levels of expenditure on different activities and assess if these are within the range for which any effect can realistically be expected;

- better, they should be calculated systematically with data from pilot studies and operational programs, and related to expected effect for example as shown in Figure I.1.

Evaluating Targetting Objectives

The target populations of the program should be defined in the plan by their numbers, characteristics, etc. The plan must also provide for adequate means of identifying these groups; if for example "malnourished children" were the stated target group, a procedure for screening and admitting them to the program must be planned for. If an estimate of the prevalence of malnutrition at the beginning of the program is available, the indicators in Part A of Table I.8 should be calculated. If such an estimate is not available and cannot be approximated to, these indicators cannot be calculated. For planning purposes, the appropriate indicators are the proportion of targetted who are needy (planned focussing), and the proportion of needy who are targetted (planned coverage). This provides the basis for comparing planned with actual focussing and coverage. Estimates of probable effects can be obtained if these indicators can be calculated from pre-program data, using only data on program participants, as discussed in Mason et al. (1982) Chapter IV. Apart from their use to evaluate the actual implementation of a program, the indicators allow evaluation of the program plan itself. For example, if focussing is not greater than the population prevalence of needy, the program is not targetted operationally to the needy (for instance, a program that was not targetted at all, that is covered the whole population of an area, would have a value of focussing the same as the population prevalence). Judgements on how far the program plan was actually oriented to the malnourished, and on the relative efficiency of different targetting strategies, may be useful later in making recommendations on modifications to the program.

The calculations can be extended further by putting in costs. For example, if there are differential costs involved in reaching alternative target groups, knowledge of their relative nutritional status allows calculation of the optimum combination of delivery to different groups to minimize cost per malnourished reached.

Evaluating Outcome Objectives

The distinction has been made between the population as a whole, the (intended) target group(s), and program recipients. The expected outcome per head is reduced going from recipients to target groups to population. In

evaluating the plan, it needs to be decided (if it was not already decided in planning) to which population category the objectives apply. This may be done in one of two ways: an overall objective could be stated as a reduction of, say, malnutrition in the population as a whole, and targetting used to focus resources on groups with a high prevalence to improve efficiency; or the objective could be stated as specifically to improve nutrition in the target group.

This stage is going to mainly involve back-of-an-envelope calculations, but these nevertheless are important since they help decide what effects to look for and when they are adequate both of themselves and with due account taken of costs. At an extreme, if abolition of the problem is expected, this will be much easier to detect than e.g. a 10% reduction in its prevalence. Clearly, it is necessary to start by assessing expected effects on recipients, and then to adjust these for "planned" or "unavoidable" under-implementation and/or leakage to non-target groups. Numbers as well as prevalences should be used. The target group and the population should be the units of evaluation. The recipients should also be considered, to calculate the expenditure per caput of recipients: this will allow some estimate as to whether an adequate effect on the recipients (whether or not these are the same as the targetted) can be expected.

There is no widely accepted method for setting expected and adequate levels of outcome. However some suggestions can be made, as follows.

(1) A minimum level of expected effects on the targetted that is regarded as worth the effort, or cost, could be set. This minimum identifies the level of outcome below which the program is unsatisfactory. This minimum level of adequacy can be set as, e.g. "cases improved per dollars spent". If the program does not then meet this minimum adequacy level it should either be reexamined carefully or discontinued. This is a choice which must have been considered in advance as part of setting objectives. This assessment has to be made at some stage, and it is better to reach preliminary conclusions as early as possible, since if such conclusions cannot be reached, results from data collection will not be interpretable anyway: it is preferable to face this dilemma before embarking on data collection and analysis than to find out later that the results cannot be used for making decisions on the project. Defining this minimum level is also needed to design data collection and analysis, by defining what levels of change need to be detected.

(2) Experience from other programs could be used, particularly pilot projects, to arrive at some statement of the expected levels of

effects. As discussed in Part I, data from recent reviews of 21 nutrition and health projects (Habicht & Butz 1979; Gwatkin et al. 1980; Drake et al., 1980) show some consistency in results. These projects seem to have achieved, for example, a reduction in the prevalence of malnutrition of around 5 to 15 cases per 100 for a baseline prevalence of 30%, or an improvement of about 2 to 6% in mean weight for age. All of these projects were pilot projects. The cost estimates for implementation at pilot level (which may differ considerably from estimates for scaled up operations) are around $1 - $8 per head per year. The time periods vary, as does whether the effects are on the entire target population or on the specific individuals who fully participated in the projects. With a per capita cost of $5 per head per year, and a reduction in prevalence of malnutrition of 10%, the effect per cost calculates out at 20 cases of malnutrition rehabilitated or prevented per $1,000 expenditure. (The distinction between cost per head of population (e.g. target population) and costs by cases prevented or rehabilitated is obviously important, these being related by the program effectiveness). Useful values for effects per unit cost in terms of cases prevented or rehabilitated are given in Beghin (1980).

(3) Certain biological relationships are approximately known: if dietary intake is taken to 100 or more percent of requirement, after, say, one year any child actually eating this amount of food should be no longer malnourished. By the same token, expected impacts on health from control of infectious diseases are not totally unknown.

One purpose of the above exercise is to eliminate the need to go further in the evaluation of some projects - there may be many - which could not be expected to have enough of an effect to warrant their continuation, much less their expansion; because the inputs could not produce the hoped-for outcome, because of inadequate targetting, and so on.

At this stage, relevant outcome indicators should be tentatively identified. The starting point is clearly the outcome objective. If decreasing the prevalence of protein-energy malnutrition in pre-school children is an objective, suitable indicators of nutritional status in these children will be required, the choice depending primarily on data availability and responsiveness of indicators to the intervention. If improved school performance is an objective, data on this

performance will be required. Table I.10 has given some information on indicators by the type of program and its objectives. They are further discussed under Stage 4.

For those programs that survive this reappraisal, the process provides a clue as to how easy or difficult it will be to detect an effect, and how large that effect will have to be to be considered adequate. Our bet is that such calculations will show that a substantial number of projects require no further evaluation, or at least merit little further expenditure on seeking impacts that are at best hardly important to the overall problem. Several (hypothetical) examples may be cited, as discussed below.

- School feeding programs are an obvious example; their objectives are all too frequently mis-directed. These objectives may be specified by the numbers of school children affected, the number of schools, the levels of food provided, etc. "Reduction in malnutrition" is often specified in these programs as the planned outcome. But these programs are targetted at an age-group generally with a low prevalence of malnutrition; they reach, in many countries, only those children who are better-off and can, hence, enroll in schools; and finally they often emphasize foods of low priority - e.g. protein when the deficit in the diet is primarily of total food. Thus an objective stated as substantially decreasing malnutrition is often likely to be shown to be unrealistic by examining the work plan.

- Pre-school supplementary feeding programs, on the other hand, are aimed at the age group most affected by malnutrition. A typical example might be the following: all the malnourished children attending clinics in an area are to be provided with 500 kcals per day for three months. Suppose the area contains 1,000,000 people of whom 200,000 are under five years of age. The expenditure is $500,000 per year. 30% of the children are malnourished, that is 60,000 and 10% of these attend clinics, i.e. 6,000. The questions which immediately come to mind when evaluation of the program is being considered are:

  - Are the children who receive food in fact malnourished?
  - Do these children improve?

- Does the program have a substantial effect on the overall problem (the answer to this is clear, since 90% of the needy children do not attend clinics).

If 50% of the needy children receiving food were rehabilitated, this would mean that 3,000 children were improved, at a cost of $500,000. This gives an effect per unit cost of 6 rehabilitated per $1,000. It might be decided in planning that this was just acceptable. If so it means that if an evaluation found that less than 50% of those malnourished children attending clinics were rehabilitated, or if less than 10% of the malnourished attend clinics at all, the program should be carefully reconsidered, because the effect per unit cost will be below 6 rehabilitated per $1000.

o Another example, comparable with the previous one, might be a targetted food distribution program, with eligibility criteria. Again consider a million people, 200,000 under 5's, an expenditure of $500,000 per year, 30% (60,000) of the under 5's malnourished. If the program reaches all these children, the expenditure would be around $8 per head per year. For a typical family, out of a total expenditure of say $300 per year, some $200 may be spent on food, i.e., perhaps $30 per year for a child. Hence the $8 per child head per year provided as food is quite significant, and it might be expected that a reduction from 30% malnutrition rate to, say, 20% in the population could reasonably be expected. This implies that, of the 60,000 children initially malnourished, 20,000 are no longer malnourished. Hence for each $1000 of expenditure 40 children are rehabilitated. In this case, if a change of prevalence from 30% to 20% was indeed detected, there might be good reason to suppose that we are looking at an effective program. Hence, further investigation might be in order to gain more confidence that this improvement was indeed due to the program. Similarly, an effect per unit cost of 20 cases prevented per $1000 might be regarded as just adequate, so that the evaluation would be designed to determine whether the prevalence was now above or below 25%.

o A rehabilitation clinic might cost $10,000 per year to run and succeed in rehabilitating 100 cases of malnutrition. This gives a

TABLE II.3

STAGE 3: EVALUATING IMPLEMENTATION

1. **Does program reach intended target group?**

   Assess targetting as:

   - Numbers targetted who are recipients, hence calculate indicators of delivery and leakage (Table I.8)

   - If outcome data available, (i.e. number needy who are recipients), calculate indicators of actual focusing and coverage (Table I.8), compare with targetting objectives.

2. **Assess level of delivery as:**

   - Expenditure or other measure per caput of recipients

   - Compare with implementation objectives.

3. **Do deviations from objectives affect expected outcome** (for target group or population)?

   - If to point where adequate effect could not be reached - **stop**

4. **How should implementation be improved?**

   - Reassess targetting, delivery (e.g. expenditure per caput)

figure of 10 cases per $1,000. However, even if this effect per unit cost is acceptable, there will usually be logistical and personnel reasons for not scaling up such a project so that its coverage could never be satisfactory. Therefore even if this was fully effective, it may not a priori be justified; hence outcome evaluation is not necessary, and evaluation of the plan alone will suffice.

- In famine relief the only issue really is whether severe malnutrition remains in the location (a camp, or an area) after famine relief has been going on for some time. The role of food in improving the situation probably requires no elaborate assessment -evidently food is needed and there is unlikely to be any question of withholding this on the grounds that its effect cannot be shown. An evaluation here is likely to focus on procedures, in order to gather lessons for future, more efficient implementation of famine relief. Again, no detailed analysis of cost-effectiveness or attempts to attribute changes in outcome to food inputs are likely to be appropriate for administrative or managerial purposes. There may still be a research need to establish causality in the effect of famine relief, although this is difficult to design and will not be addressed further here.

STAGE 3: EVALUATING IMPLEMENTATION

(See Table II.3)

Evaluating the implementation of a program is reasonably straight-forward in principle. It involves the following specific questions.

1. Have project components been delivered as expected to the intended target groups?
   - Was the selection of recipients correct?
   - Was expenditure or other measure of delivery per caput of recipients adequate?
   - Was coverage adequate?
2. Have deviations in project implementation been such that, as a result, outcome objectives could not be reached?
3. Can constraints to implementation be identified? How can they be removed to improve program delivery? How should the plan be changed?

Targetting and delivery

The aim of evaluating targetting could be summarized as filling in Part B of Table I.8. This then allows derivation of indicators of delivery and leakage, meaning proportion of targetted who are recipients, and proportion of recipients not targetted, respectively.

Quantitative assessment of level of delivery e.g. expenditure per caput must rely on administrative records; retrospective interviews with participants (e.g. by sample survey) cannot fully replace these records. Relevant records pertain to financial aspects, delivery of goods, staffing, etc. This stage of the study therefore uses administrative records to:

- Examine whether delivery is according to plan
- Reassess delivery in order to update estimates of likely effects (e.g. to see if expenditure per recipient is still within the range in which an effect can be expected).

Assessment of the extent to which the program has been delivered to the needy, and to which the needy are covered by the program, requires outcome data but logically are included here as it concerns implementation rather than effect of the program. If these ratios can be calculated from available data, they will allow important estimates to be made of the success of the program's implementation. For example, focusing is directly related to effect on the target group, per unit cost (not only on recipients). This is equivalent to filling in Part C of Table I.8.

Again, a priori decisions may be needed as to how far deviations from planned targetting should be acceptable. In any event, these values will show how the actual delivery to the target group needs to be modified. The question of whether the right people are reached is measured by focusing; whether enough are reached is measured by coverage.

This examination and reassessment relies on process indicators. Some examples relevant to nutrition/health programs might be the following: timing of delivery of supplies and equipment; participation by attendance or by receipt of food, immunizations, health care, health and nutrition education; staff performance measured by the number of contacts with recipients per worker, or by the number of contract hours per worker; etc.

Improving Implementation

There are two purposes to these analyses. First, to see whether the program delivery is sufficiently in line with the plan that an adequate outcome

can still be expected. This should be based on the premises used to design the program, maybe modified from more recent knowledge.

The second purpose, of important concern in practice, is to allow identification of constraints and failures in implementation. This is both in terms of degree of implementation (e.g. quantities of goods and services), and targetting. Recommendations from these results may be among the most valuable outputs of the evaluation, certainly for the program's management, and for its administration and funders.

STAGE 4: EVALUATING GROSS OUTCOME

(See Table II.4)

The evaluation of the adequacy of program effects relies in the first instance on assessment of gross outcome - that is on detected change in outcome indicators, not allowing for the effect that might have occurred anyway. This involves comparison of the gross outcome with a pre-established standard of adequacy of effects. By definition the gross outcome of a program refers to the absolute or relative change observed in one or more indicators of program effects. The evaluation of gross outcome has to cope with three issues:

(a) How are outcome indicators to be chosen?

(b) How are they to be measured?

(c) How are they to be evaluated?

Choosing Outcome Indicators

The choice of outcome indicators involves a number of considerations (specific examples are given in the relevant chapters elsewhere in this book):

- responsiveness of the indicator to the intervention (see Table I.9)
- feasibility and cost of collecting and interpreting the necessary data;
- whether the indicator is a direct or an indirect (proxy) measure of the achievement of the objective — direct measures are usually preferable.

In Table I.9 the known responsiveness of certain indicators has been shown, and a more comprehensive if less well-established list of possible indicators has been give in Table I.10.

Interviews and observations by an ethnographer or someone with similar skills should precede any systematic, quantitative data collection. Such interviews will often permit restricting quantitative data collection to a few variables and permit simpler sampling than if one tries to cover every

TABLE II.4

STAGE 4: EVALUATING GROSS OUTCOME

1. **Choose outcome indicators.** Consider:

   - Responsiveness (Table I.9) and suitability (Table I.10)

   - Feasibility and cost of collection

2. **Measure gross outcome.** If data only obtainable on program participants, try:

   - Time-in-program method

   - Rapid collection of time-series data

   - Correlational analysis

   If data obtainable on non-participants and confounding variables see Stage 5, Table III.5.

3. **Evaluate gross outcome**

   - Is it adequate? (do outcome indicators meet preset standards of adequacy? - see Stage 2 point 4) - if so, usually **stop** for routine evaluation.

   - Is there reason to believe that further data collection or analysis would increase the estimated effect (e.g. evidence for negative confounding): if not, **stop.** If so, and answer worth the cost, continue to assess net outcome (Stage 5)

eventuality. This can result in major savings in time and money, especially if the program is inadequately implemented so that no quantitative collection of outcome data is justified.

Measuring Gross Outcome

The measurement of gross outcome does not always preclude subsequently moving by analysis to assessing net outcome. This depends both on the design of the evaluation and whether data on confounding variables are collected for later analysis (see Table I.5).

Measurement of gross outcome preferably entails collection of baseline data from program subjects before the program begins, and a similar effort at a point in its operation sufficiently later to allow the program to have some measurable effect. Such data however may not always be readily available. Baseline data on program subjects are often lacking, possibly because an evaluation of the program effects may not have been the intention at the beginning. Furthermore the collection of such data may involve considerable cost if for example an extensive sample survey is required at two points in time. Certain conventional designs have been given in Part I (see Table I.6). We have considered below the special case of when only data on program participants are available, without baseline data. It is possible to rely on routinely collected program data to get approximate measures of gross outcome. Under these restrictive but common conditions three possible designs come to mind, using data only from program contacts:

(a) The time-in-program method

(b) Rapid collection of time-series data

(c) Using variations in socio-economic status and program delivery, which can be cross-tabulated or correlated

(a) **The time-in-program method**

This means comparing outcome information for children who have been in a program for a substantial period with others just entering, on a cross-sectional basis. An improvement in those who have been "treated" for a certain time over those just entering can then lead to an estimate of gross outcome. Finally, comparisons should be made between children of the same age to exclude the maturational effect. One obvious confounding factor is self-selection - e.g. those that enter later have different nutritional status on entry than those admitted to the program at the beginning. This can be checked if nutritional status on entry into the program is recorded, as is often the case. Another

problem is regression to the mean - if malnourished children are selected, they may be an extreme that anyway would improve. Nonetheless, with commonsense allowance for such factors, this procedure provides some internal control.

(b) **Rapid collection of time-series data.**

The effects of certain programs may in fact be detectable by time-series observations within short periods. The time for carrying out and analyzing a cross-sectional survey is itself going to be around six months at least - time enough for a feeding program to show its effects on nutritional status of participants. This can at least serve as a check that the program has an effect on those to whom it is delivered. The question then reverts to whether these are the target group, etc. Examining these latter questions may, of course, require further data.

(c) **Correlation studies: using variations in program delivery and socio-economic status on a cross-sectional basis.**

If variations in both program delivery and socio-economic status can be measured for program participants, then certain useful conclusions can be reached. The variation in program delivery is likely to be most obvious between different sites. In practice, in certain locations the program is going to be more efficiently delivered than in others, and an important question becomes how to measure this. At the same time, it is most likely that socio-economic status will influence nutritional status, and it may not be too difficult to postulate certain summary indicators of this. For example, quality of housing has been shown to be associated with nutritional status. If this can be done, then, in its simplest form, a limited number of possibilities may be set up, using nutritional status of program participants; these have been discussed in Part I, see Table I.7.

Evaluating gross outcome essentially relies on a limited number of variables pertaining to the nutritional status (weight, height, age, etc. which may be indicators themselves or from which indicators may be derived), the extent of delivery of the program services and the duration of time that a participant benefits from a program. Such information is used to establish relationships between program variables on the one hand and measures of gross outcome on the other. For example, a positive correlation between the extent of delivery and weight-for-age of the participants may indicate that a longer period of participation is associated with a more adequate outcome. This may be taken as an indication that the gross outcome of the program has been positive. Where

the measure of nutritional status of a participant is also available before he joins the program, a comparison of this information with a similar measure after some length of time also measures the gross outcome if maturational effects and regression toward the mean are to be taken into account.

Finally, comparison of estimated gross outcome with program costs will allow at least a rough estimate of whether the effects per unit cost detected so far are in the range likely to be regarded as adequate. If they are far below the adequacy level, it may not be necessary to estimate net outcome at all, since it may only further decrease the estimated effect per unit cost, or at least may not dramatically improve it. Many of the confounding factors discussed in Table I.5 usually act to give over- estimates of the net outcome (positive confounding). With an inadequate gross effect per unit cost, it may be worth considering whether more accurate assessments are in fact likely to increase this value - i.e. there has been negative confounding. (e.g. Chernichovsky 1979). It is our impression that the more careful the evaluation study, the less effect on outcome is usually found. Therefore, unless there is good reason to believe that without-program values would have deteriorated sharply, all the necessary information for outcome evaluation may now be available, and it is unnecessary to go further in data collection for analysis.

STAGE 5: EVALUATING NET OUTCOME

(See Table II.5)

The sequence of stages discussed so far precedes the evaluation of net outcome since these preceding stages might lead to an early decision that further evaluation is unnecessary. This would be the case if the objectives of the evaluation have already been accomplished, for example if the program has been shown to have been inadequately implemented (e.g. in terms of delivering the intended amounts of goods and services to the targetted population). On other occasions, the program objectives may be judged as unattainable by examining the program plan. The decision on whether or not to proceed to evaluate the net outcome of a program depends on whether or not it is necessary to do so and whether or not it can be done.

At this stage, an estimate of the gross outcome will be available. If further information as to how far observed changes are likely to be due to the program is needed (i.e. moving from gross to net outcome) there are several possibilities. These have been introduced in Part I.

TABLE II.5

STAGE 5: EVALUATING NET OUTCOME

A. If data estimating gross outcome are available:

- Use data on (i) varying program delivery
  (ii) possible confounding factors (e.g. socio-economic status)

  to:
  - cross-tabulate by these factors and compare groups
  - control statistically for confounding and correlate with program delivery using multiple regression

This may give more plausible inferences on association of program with outcome (i.e. estimate net outcome). If these data are **not** available, or more certainty is needed, and surveys are considered worthwhile, go to B. If not **stop.**

B. If survey to be carried out:

- Decide on design (see Table I.5)
- Examples are:
  - interrupted time- series
  - non-equivalent control groups
- Analyze by:
  - matching
  - cross tabulation (as A)
  - control statistically and correlate with program delivery analogously to A

This may give additional plausibility to inferences on association of program with outcome.

C. Evaluate estimates of net outcome

- Is net outcome adequate?
- Is effect per unit cost adequate?
- How should program be modified (e.g. targetting, delivery, design)

(a) If there is information on varying levels of program delivery, associations of levels of delivery with outcome can be investigated statistically.

(b) If, as well as this, there are data on possible factors confounding the relationship between program delivery and outcome (e.g. socio-economic status), then these confounding variables can also be taken into account.

In the first place, this can be done by cross-tabulating, e.g. for varying program delivery and socio-economic status — see example in Table I.7. This may also give insights on targetting. Secondly — and particularly if a number of possible confounding variables are measured — multiple regression analyses (correlational analyses) can be used to simultaneously take account of several confounding variables as well as program delivery. In this case, one is investigating the effect of program delivery controlling for, e.g. socio-economic status: the sign, significance, and magnitude of the coefficients for variables representing program delivery are estimated in relation to outcome, with confounding variables in the equation.

The next step would be to control also by design, see Table I.6. This may sometimes be done by retrospective matching, but is more effective when designed into the evaluation before the intervention is initiated. The analyses are similar to those discussed above - tabulation by varying program delivery (including with/without program, before/after, etc.) and by confounding variable e.g. socio-economic status; and simultaneously taking into account these variables in the multiple regression. The plausibility of the conclusions depends substantially on the choice of design, and can be increased if the effects of confounding variables are properly controlled for either by design or analysis or both. Two possible designs, which have been referred to in Part I and which may be suitable for such situations (see Casley & Lury 1982) are described below. (See also Table I.6)

Interrupted Time-Series design

When a without-program comparison group is not available, comparisons can be made with outcome data on the treatment group before the treatment begins. This usually requires a series of observations before and after treatment. If the effect of the treatment is sufficiently "sharp" - giving a rapid change in values of the outcome variables and this may provide evidence for an effect of the program.

This design is not always possible for the evaluation of nutritional effects of programs since time-series data on the nutritional status of a population before the program may not exist and post-intervention collection of time-series data may require a long period of time and be costly. If such data were, however, available regression techniques can be employed to answer the question of whether the intervention had changed the trend of changes in nutritional status of the population in a significant way and the direction of this change. In this case, a variable representing time is introduced into the equation, and a variable representing program delivery is analyzed with respect to outcome, again taking account of other confounding variables.

Non-equivalent Control Group Design

When having a comparison group and a pretest is feasible, but randomization is not, a non-equivalent control group design may be used (see Table I.6). Likely confounding variables should first be identified, measured, and controlled for statistically in the analysis. Analysis can be done through matching or statistical control or both.

Matching seeks to identify the major confounding variables and constructs treatment and comparison groups such that they would resemble each other as closely as possible on the matching criteria. Matching can be attempted either before the treatment begins or after it has ended and the outcomes have been measured. The success of this approach depends on the extent to which the confounding variables correlated with the outcome can be identified. The difficulty with matching is that, first, it is not always evident what these major confounding variables are and second, there are likely to be quite a number of them which complicates the matching process.

Statistical control to estimate net outcome tries to isolate the impact of the intervention by compensating for the differences that may exist between the treatment and comparison groups. The methods include simple techniques of standardization and stratification and correlational analyses. All of these statistical methods are particular applications of various regression techniques (see Judd and Kenny 1981 and Anderson et al. 1980).

Much of the data required will be derived from sample surveys. The variables include indicators of outcome (e.g. nutritional status) and major potential confounding variables that can be identified (e.g. socio-economic status). Time-series data are preferable to cross-sectional data, although more often only the latter can feasibly be collected. Some considerations in designing

the survey include:

- Sampling:
  - (a) The sample should, if feasible be randomly selected, so that the probability of inclusion of each household or individual in the sample is known. This allows inferences to be made concerning the population from which the sample is drawn. Considerations for deciding sample size are given in Part I.
  - (b) Stratification: participants and non-participants should be sampled separately if they are clearly distinguishable. If not, post-stratification is used. Major potential confounding variables can also be used for stratification - e.g. geographically, by socio-economic status.
  - (c) Confounding variables: major potential confounding variables are selected on the basis of past investigations and a priori knowledge. Examples are: socio-economic status, e.g. by income, wealth indicators; community variables such as isolation, access to services, ecological conditions, water supply and sanitation, participation in other programs, etc.

A large body of literature exists on such methods for social program evaluation. Many of these have been quoted here and in Part I. These evaluations are generally in the nature of evaluative research rather than operational program evaluations, and will not be further expanded upon here.

STAGE 6: DATA PRESENTATION AND DISTRIBUTION OF RESULTS

(See Table II.6)

When the analytic work is completed, both qualitative and quantitative results must be presented as simply and didactically as possible, without however, omitting important information or considerations. This part of the report because of its quantitative nature and explication of analytic and epidemiological considerations will be accessible above all to professionals with appropriate training. This audience is important because these professionals give the report credibility. They are, however, rarely the decision-makers. The decision-makers are usually senior program managers or politicians, who need an "executive summary" of the findings. This precedes the document and should

TABLE II.6

STAGE 6: DATA PRESENTATION AND DISTRIBUTION

To whom?

- Already prepared principals (manager or funder) should have been involved in draft of report

How?

- Often iterative presentations to improve usefulness of report by better addressing needs of recipient both as to what is addressed and how (requires well done Step 1 in temporal organization of evaluation)

Feedback to other than principals

- Do it with principals; ethics of confidentiality and censorship.
- busy people need short summaries of decision options + results + recommendations; careful people need everything relevant --fill these two needs separately.

never be longer than two pages. It clearly presents what decisions are needed given the report's findings and presents only those findings necessary for these decisions.

STAGE 7: ON-GOING EVALUATION

Many of the difficulties discussed here may be mitigated by building an evaluation procedure into the implementation of the project. In some cases a decision to do this could be an important outcome of the sort of evaluation described here, which will often be "mid-term". Some advantages of doing this are:

- baseline data from the mid-term evaluation will be available;
- time-series data can be organized;
- adaptation of the program can be much more usefully made as soon as the need becomes apparent (rather than complaining about what should have been done some time previously as in many post facto evaluations).

The most basic approach to on-going evaluation has been set out in the context of nutritional surveillance, which we have referred to as "adequacy evaluation" (see Mason et al. 1982). This is intended mainly for the use of program managers. Adequacy evaluation covers both process and outcome. It essentially addresses two questions:

(a) Is the program being delivered as planned to the intended target group? (i.e. the same questions as for process evaluation)

(b) Is the (gross) outcome acceptable?

This second question could be elaborated as:

- is the trend in outcome indicators adequate for the program target groups? or in the population as a whole?

Answers to both these questions lead directly to decisions on program implementation. A negative answer to the first question leads to re-examination of the program organization and management. A negative answer to the second question should lead to further investigation as to why the program is apparently failing to meet its objectives in terms of effects on the population.

There are two requirements for this adequacy evaluation, similar to those for one-time evaluations. First, clear and quantified definition of target groups is needed. Second, a definition of adequacy; this involves both defining the units in which outcome is to be measured, and setting levels of these units, a deviation from which will be considered inadequate and will lead to further action. It is

important to note that from adequacy evaluation, it should be possible to derive costs relative to activity, and costs relative to gross outcome, either for the program target group or in the population as a whole. It will not be possible to derive a true estimate of cost-effectiveness, meaning cost per unit of net outcome due to the project, because only gross outcome is assessed.

There are few examples from <u>any</u> field at the present time of successful continuing evaluation of outcome; and even good monitoring of process is not common. Nonetheless, experience is being gained, and serious attempts to put forward and apply practical approaches are being made. For example, a useful handbook on monitoring and evaluation of agricultural and rural development projects has recently been published by the World Bank (Casley & Lury 1982), and steps have been taken to use methods such as these (which include nutrition as an outcome variable to evaluate development projects). Partly progress depends on allocation of the necessary funds, and the political will actually to evaluate. As noted in Part I, around 0.5 to 5% of project costs have been allocated to evaluation, with 1% the norm for health programs in the U.S. Similar allocations to food aid and nutrition programs would probably buy useful evaluation systems.

The state-of-the-art is not as advanced as we might like, but it is probably adequate to allow on-going evaluation for adequacy or impact assessment to be usefully built into programs much more frequently than at present.

References

Anderson, S.; Auquier, A.; Hauck, W.W.; Oakes, D.; Vandaele, W.; and Weisberg, H.I. 1980. Statistical Methods for Comparative Studies. Techniques for Bias Reduction New York: John Wiley & Sons.

Beghin, I./FAO 1980. Selection of specific nutritional components for agricultural and rural development projects. Mimeographed. Antwerp, Belgium: Nutrition Unit, Institute of Tropical Medicine.

Casley, D. and Lury, D. 1982. A Handbook on Monitoring and Evaluation of Agricultural and Rural Development Projects. Baltimore: Johns Hopkins Press.

Chernichovsky, D. 1979. The economic theory of the household and impact measurement of nutrition and related health programs. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et al., pp. 227-67. New York: Plenum Press.

Drake, W.D.; Miller, R.I.: and Humphrey, M. 1980. Final Report: Analysis of Community-Level Nutrition Programs. Project on Analysis of Community Level Nutrition Programs. Vol. I. Washington, D.C.: USAID - Office of Nutrition.

Gwatkin, D.R.; Wilcox, J.R.; and Wray, J.D. 1980. Can health and nutrition interventions make a difference? Overseas Development Council Monograph No. 13. Washington, D.C.: ODC.

Habicht, J-P. and Butz, W.P. 1979. Measurement of health and nutrition effects of large scale nutrition intervention projects. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et. al., pp. 133-82. New York: Plenum Press.

Judd, C.M. and Kenny, D.A. 1981. Estimating the effects of social interventions. Cambridge: Cambridge University Press.

Mason, J.B.; Habicht, J-P.; Tabatabai, H.; and Valverde, V. 1982. Nutritional Surveillance. Forthcoming - presently mimeo. Ithaca, New York: Cornell University.

Riecken, H.W. 1979. Practice and Problems of Evaluation: A Conference Synthesis. In Evaluating the Impact of Nutrition and Health Programs, eds. R.E. Klein, et al., pp. 363-86. New York: Plenum Press.