© International Epidemiological Association 1999 Printed in Great Britain

LEADING ARTICLE

Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact

JP Habicht,[a] CG Victora[b] and JP Vaughan[c]

The question of why to evaluate a programme is seldom discussed in the literature. The present paper argues that the answer to this question is essential for choosing an appropriate evaluation design. The discussion is centered on summative evaluations of large-scale programme effectiveness, drawing upon examples from the fields of health and nutrition but the findings may be applicable to other subject areas.

The main objective of an evaluation is to influence decisions. How complex and precise the evaluation must be depends on who the decision maker is and on what types of decisions will be taken as a consequence of the findings. Different decision makers demand not only different types of information but also vary in their requirements of how informative and precise the findings must be. Both complex and simple evaluations, however, should be equally rigorous in relating the design to the decisions. Based on the types of decisions that may be taken, a framework is proposed for deciding upon appropriate evaluation designs. Its first axis concerns the indicators of interest, whether these refer to provision or utilization of services, coverage or impact measures. The second axis refers to the type of inference to be made, whether this is a statement of adequacy, plausibility or probability.

In addition to the above framework, other factors affect the choice of an evaluation design, including the efficacy of the intervention, the field of knowledge, timing and costs. Regarding the latter, decision makers should be made aware that evaluation costs increase rapidly with complexity so that often a compromise must be reached. Examples are given of how to use the two classification axes, as well as these additional factors, for helping decision makers and evaluators translate the need for evaluation—the *why*—into the appropriate design—the *how*.

Keywords Evaluation, public health, diarrhoea, impact

Accepted 1 April 1998

Funding agencies are increasingly requiring quantitative evaluations of the impact of public health programmes, to meet increased demands for accountability. The present paper addresses summative evaluations of established interventions, rather than formative evaluations whose purpose is to fine tune programme implementation.[1] The results of summative evaluations are to be used to make decisions about the programmes evaluated. Such 'instrumental' use of evaluation results is on the increase.[2] This is a distinct situation from what was observed in the past, when evaluations had limited 'instrumental' use but affected programmes and policies less directly, through changing perceptions. This difference in the uses of evaluations is important because one is more likely to reach decision makers when the use is 'instrumental', since the evaluators can ascertain what information is necessary for the decision-taking.

It is generally understood that other factors weigh as much or even more than quantitative evaluation results in the final decisions about programmes. However, the inferences from quantitative evaluations should be pertinent to the decisions if

[a] Division of Nutritional Sciences, Cornell University, USA.

[b] Universidade Federal de Pelotas, Brazil.

[c] London School of Hygiene and Tropical Medicine, UK.

Reprint requests to: Prof. CG Victora, Departmento de Medicina Social, Universidade Federal de Pelotas, CP 464, 96001–970 Pelotas, RS, Brazil.

these evaluations are to have any utility, and thus experienced evaluators design their evaluations to address the specific questions of concern to decision makers.[3,4] Of great importance for evaluation design is the type of inference required by decision makers, an issue which so far has not been addressed in the epidemiological literature.

This paper uses conventional epidemiological designs to discuss the above points, drawing from the authors' experience in the fields of health and nutrition in developing countries. It is particularly addressed to assessing effectiveness, that is, the large-scale achievements of interventions which, under ideal controlled conditions, have a known efficacy.

Why do the evaluation? Who will be influenced?

It is well recognized that formative evaluations must be done with those who have authority for the changes that need to be instituted. Less well understood is that the same kind of participatory research is essential in almost all summative evaluations for the information to be actually appropriately used in decision making.

Based on the findings from a summative evaluation, a decision maker may decide to continue, change, expand or end a project or intervention. The first task for the evaluation planner, therefore, is to define the target audience for the evaluation results, since the responsibilities and expertise of the decision makers will affect what questions should be asked.

Different decision makers not only ask particular questions but also require distinct kinds of inferences from the quantitative data. In other words, the answer to the question on *why* do an evaluation will affect its inferential design. For example, a donor agency may wish to document a statistically significant impact on mortality, while a district health manager may be interested in knowing whether a certain coverage was reached if the cold chain is functional. This does not imply that one kind of evaluation is more 'scientific' than the latter, as both types can and should be equally rigorous, in the sense of providing information that is sufficiently valid and precise for the decisions to be taken. The first type of evaluation provides evidence of effectiveness, being relevant to a decision to expand the programme. The second, on the other hand, assesses the overall adequacy of changes in outcomes, and may support a decision that no changes are required.

The evaluation designer should thus work with the decision makers for planning a study that will satisfy their requirements, that is, which will address the *why*. A conceptual framework is presented below to help *how* to design the evaluation. Note that an evaluation may be aimed at more than one category of decision maker. In this case, the design must take into account their different needs.

Classification axes

The proposed classification is based on two axes. The first refers to the indicators, that is, whether one is evaluating the performance of the intervention delivery or its impact on health or behavioural indicators. The second axis refers to the type of inference to be drawn, including how confident must the decision maker be that any observed effects were in fact due to the intervention.

First axis: What do you want to measure? Indicators of provision, utilization, coverage and impact

A useful way of looking at evaluations of health and nutrition interventions is to ask what is to be evaluated. The answer to this question will determine what will be measured. One may evaluate the provision or utilization of services, coverage or impact. Table 1 presents the outcomes of interest in a logical order leading from provision to impact. The services must be provided so that they are available and accessible to the target population and of adequate quality. Second, the population must accept the services and make use of them. Third, this utilization will result in a given population coverage. Coverage is a particularly useful measure, representing the interface between service delivery (the managerial process) with the population (the epidemiological picture). Finally, the achieved coverage may lead to an impact on behaviour or health. Any important shortcomings at the early stages of this chain will result in failures in the later achievements. For each outcome Table 1 presents a relevant question and an example of an indicator useful in the evaluation of a programme for the control of diarrhoeal diseases aimed at young children with emphasis on the promotion of oral rehydration solution (ORS). In subsequent tables the term *performance* evaluation will be used to encompass evaluations of provision, utilization and coverage, as separate from *impact* evaluations.

The evaluator should choose the indicators on the basis of discussions with the decision makers. The complexity of the evaluation designs and the extent of data collection will also depend on the decision maker's intended use of the results. As discussed, local managers may need summative data on provision and utilization to improve them within a health centre or in a district. On the other hand, national or international agencies may require assessments of coverage or impact to justify further investments in the programme. It also depends on how much one is willing to pay for the evaluation. Provision or utilization may be assessed by visiting services or using routine information systems. Coverage or impact, however, almost always require field data collection with important cost implications.

Second axis: How sure do you want to be? Types of inference: adequacy, plausibility, probability

The second axis refers to the kind of inference (adequacy, plausibility or probability), as well as on how confident decision makers need to be that any observed effects are in fact due to the project or programme. Both performance and impact evaluations may include adequacy, plausibility or probability assessments.

Adequacy assessment: Did the expected changes occur?

Inferences about the adequacy of programme outcomes depend on the comparison of the performance or impact of the project with previously established adequacy criteria. These criteria may be absolute—for example, distributing 10 million packets of ORS to children with diarrhoea or achieving 80% ORT use rate—or may refer to a change—for example, a 20% decline in reported diarrhoeal deaths in the programme area. Even when specific goals have not been established, performance or impact may still be assessed by measuring general time trends, such as an increase in coverage or a reduction in mortality.

Table 1 Example of indicators for evaluating a diarrhoeal diseases control programme

Indicator	Question	Example of indicators
Provision	Are the services available?	• No. of health facilities offering CDD activities per 100 000 population
	Are they accessible?	• Proportion of the population <10 km of a health facility with CDD activities
	Is their quality adequate?	• Proportion of health staff with recent CDD training
Utilization	Are the services being used?	• No. of attendances of under fives with diarrhoea per 1000 children
		• No. of ORS packets distributed
Coverage	Is the target population being reached?	• Proportion of all under fives with diarrhoea who used ORT
Impact	Were there improvements in disease patterns or health related behaviours?	• Time trends in diarrhoeal deaths and hospital admissions

Table 2 Characteristics of adequacy evaluations

Type of evaluation	Measurements	In whom?	Compared to what?	Inferences
Adequacy			Predefined adequacy criteria	Objectives met
Performance (**provision, utilization, coverage)**	Programme activities	Implementation workers Programme recipients		Activities being performed as planned in the initial implementation schedule
Cross-sectional	Once		Absolute value	
Longitudinal	Change		Absolute and incremental value	
Impact	Health and behavioural indicators	Programme recipients or target population		Observed change in health or behaviour is of expected direction and magnitude
Cross-sectional	Once		Absolute value	
Longitudinal	Change		Absolute and incremental value	

Adequacy assessments require no control groups if results are to be compared with set criteria (e.g. 90% exclusive breastfeeding rate by the age of 4 months). For assessing the adequacy of change over time, at least two measurements will be required, thus increasing the complexity of the design. Nevertheless, adequacy assessments are usually much less expensive than the other two types.

The main characteristics of adequacy evaluations are summarized in Table 2. Adequacy performance evaluations assess how well the programme activities have met the expected objectives. For example, these may include assessments of how many health centres have been opened, how many ORS packets or other drugs are available, how well health workers have been trained, how many children used the services or what coverage has been achieved in the target population. The evaluation may be cross-sectional, carried out on a single occasion, during or at the end of the programme. It may also be longitudinal, requiring baseline data or including repeated measurements for detecting trends.

Adequacy impact evaluations assess whether health or behavioural indicators have improved among programme recipients or among the target population as a whole. Again, the assessment may be cross-sectional or longitudinal. An advantage of adequacy assessments is that they can often use secondary data so that evaluation costs are much reduced.

Adequacy evaluations are limited to describing whether or not the expected changes have taken place. When assessing provision or utilization, one may reasonably ascribe an observed success to the programme being evaluated. For example, improved case management skills among health workers and increased distribution of ORS may be safely attributed to a Control of Diarrhoeal Diseases (CDD) programme. When measuring coverage or impact, however, it may be difficult to infer that any observed improvements were due to the programme since there is no control group to ensure that these changes would not take place anyway. The observed improvements may have been caused by outside influences such as secular trends in mortality or malnutrition, general socioeconomic improvements, and the presence of other projects in the same area, etc.

Adequacy evaluations may also show a lack of change in the indicators. Under usual conditions, this suggests that the programme has not been effective. However, under special circumstances—such as a general deterioration in socioeconomic situation, a famine or another emergency, or general failure of other services—a lack of change may show that the programme has been effective in providing a safety net for the affected population. This scenario is further discussed in the plausibility section.

Despite their inability to causally link programme activities to observed changes, adequacy evaluations may provide all the reassurance necessary that the expected goals are being met and lead to continued support for the programme. For many decision makers, more complex evaluation designs will not be required, particularly since these would demand additional time, resources and expertise. If the evaluation finds that the programme goals are not achieved, further evaluations may be required to identify the causes for the failure and to guide remedial action. For other types of decisions, adequacy statements must be combined with either plausibility or probability assessments to deliver the necessary inferences.

Plausibility assessment: Did the programme seem to have an effect above and beyond other external influences?

Some decision makers may require a greater degree of confidence that any observed changes were in fact due to the programme. Plausibility appraisals go beyond adequacy assessments by trying to rule out external factors—called hereafter 'confounding factors'—which might have caused the observed effects. A statement is plausible if it is 'apparently true or reasonable, winning assent, a plausible explanation'.[5] Table 3 summarizes the main types of plausibility evaluations.

Plausibility assessments attempt to control for the influence of confounding factors by choosing control groups before an evaluation is begun, or afterwards during the analyses of the data.

There are several alternatives for choosing a control group but the final choice is often dictated by opportunistic criteria, that is, by taking the best advantage of the existing situation. Control groups may include:

(a) *Historical control group*: the same target institutions or population. This approach entails a comparison of change from before to after the programme, accompanied by an attempt to rule out external factors.

(b) *Internal control group*: institutions, geographical areas or individuals that should have received the full intervention but did not, either because they could not or refused to be reached by the programme. Often, reception of a programme is variable. The indicators may then be compared between three or more groups of communities or individuals with different intensities of exposure to the intervention. A dose-response relation between intensity of the intervention and the observed performance or impact allows a stronger plausibility statement than findings from comparison between all and nothing groups. These approaches require comparisons of cross-sectional data collected at the end of the programme cycle.

Another kind of internal impact assessment is the use of the case-control method[6] to compare previous exposure to the programme between individuals with and without the disease. An advantage of the 'case-control' method is that it can be initiated relatively early after the initiation of the programme and may deliver definitive results earlier.

(c) *External control group*: one or more institutions or geographical areas without the programme. In this case, the comparison may be cross-sectional (intervention versus control at the end of the programme cycle) or longitudinal-control (comparing intervention and control at the beginning and at the end of the cycle).

The use of any of the above control groups results in much more plausible conclusions than if no controls are used. Plausibility is often markedly improved if they are used in combination. For instance, staggered interventions that begin at different times in separate areas allow the combination of historical data with external controls represented by areas where the intervention will start later; that in turn will have historical controls.

The intervention and control groups are supposed to be similar in all relevant characteristics except exposure to the intervention. This is almost never true since one of the comparison groups can be influenced by a confounding factor that does not affect the other group as much. For example, if the (control of diarrhoeal disease) CDD programme is implemented in an area with a better water supply than the control area, a difference in diarrhoeal mortality may be due to improved water and not to the programme. Dealing with confounding requires the measurement of probable confounders and their statistical treatment through matching, standardization, stratification, or other forms of multivariate analysis.[7]

Control of confounding is particularly important when internal comparisons are being made. Individuals who refuse the intervention or those who could not be reached often also differ from recipients in a number of other ways.

Confounding is also critical when using historical controls. This design is similar to an adequacy evaluation, in which a trend is recorded without external comparisons. To characterize

Table 3 Characteristics of plausibility evaluations

Type of evaluation	Measurements	In whom?	Compared to what?	Inferences
Plausibility			'Opportunistic' or non-randomized control group	The programme appears to have an effect above and beyond the impact of non-programme influences
Performance (provision, utilization, coverage)	Programme activities	Implementation workers Programme recipients (dichotomous or dose-response)		Intervention group appears to have better performance than control
Cross-sectional	Once		Control group	
Longitudinal	Change		Before-after	
Longitudinal-control	Relative change		Comparing before-after between intervention and control	
Impact	Health and behavioural indicator	Programme recipients or target population (dichotomous or dose-response)		Changes in health or behaviour appear to be more beneficial in intervention than control group
Cross-sectional	Once		Control group	
Longitudinal	Change		Before-after	
Longitudinal-control	Relative change		Comparing before-after between intervention and control	
Case-control	Once	Target population	Comparing exposure to programme in diseased (cases) and non-diseased (controls)	

a plausibility evaluation, however, one must also attempt to exclude other possible causes for the observed trends, for example, by assessing whether a decline in diarrhoeal mortality might have been due to socioeconomic development, to improved water supply and sanitation, to nutritional or other health interventions. This may be accomplished by estimating how much mortality would have decreased as a result of external changes and comparing that with the observed decline.[8] A special situation is when no important improvement was observed,[9] but using the above simulation approach one shows that a deterioration was expected. In this case, one may plausibly state that the programme was successful in preventing the situation from getting worse as a result of external hardships.

In many aspects, plausibility assessments are akin to 'natural experiments'. The evaluator will take advantage of the opportune existence of a control group to examine the effect of a programme. As its name indicates, a plausibility statement is largely based on value judgments of experts in the field, including the decision makers and the evaluators.

Plausibility assessments encompass a continuum, ranging from weak to strong statements. At the lower end of the plausibility scale are the simple comparisons with a control group, with an attempt to discuss and rule out possible confounding. At the higher end of the scale, one may have several comparisons and mathematical simulations. To reach the highest level of plausibility, one must formally discard all other likely explanations for the observed improvements. For example, plausibility would become stronger by consecutively showing that: (a) diarrhoeal mortality fell rapidly in areas with the CDD interventions (congruency of expected trend); (b) diarrhoea did not fall in the areas without the CDD interventions (not due to general changes in diarrhoea in the area); (c) changes in other known determinants of mortality could not explain the observed decline (lack of measurable confounding); (d) there was an inverse association between intensity of the intervention in the programme areas and diarrhoeal mortality (congruency of dose-response); (e) mothers with knowledge of ORT had fewer recent child deaths than those without such knowledge (congruency of mediating variables); (f) mortality among non-acceptors in the programme area was similar to that of the control area (congruency of lack of impact in the absence of the intervention); (g) the increase in ORT coverage was compatible with the degree of mortality reduction (congruency of magnitude of effect on mediating variables).

From an academic standpoint, the main shortcoming of plausibility assessments is that one cannot completely rule out all alternative explanations for the observed differences. However, by the time one had demonstrated point 'g' such alternatives are so unlikely as to be negligible. Furthermore, from a more practical, programmatic point of view, even less stringent plausibility statements are often sufficient for deciding about the future of a programme, because the cost to the decision maker of making a mistake is sufficiently low that higher plausibility is not necessary.

Probability assessment: Did the programme have an effect ($P < x\%$)?

Probability evaluations aim at ensuring that there is only a small known probability that the difference between programme and control areas were due to confounding, bias, or to chance. These evaluations require randomization of treatment and control activities to the comparison groups, being the gold standard of academic efficacy research.

While randomization does not guarantee that all confounding is eliminated (a common erroneous belief) it does ensure that the probability of confounding is measurable, being part of the error associated with the significance level used ($P < x\%$), where P is chosen on the basis of considerations discussed below under 'Magnitude of sample'. The confounding factor does not even have to be known for this procedure to work. Thus randomization assures that the statistical statement of association is directly related to the intervention. This means that the statement of statistical probability of such a 'probability' evaluation relates directly to the causality of the intervention, and is not simply a statement that the comparison groups are different as is the case for all the other designs. We will not further discuss the details of probability evaluations here since these are adequately described in standard textbooks, in particular preventing biases that accompany the intervention from clouding the evaluation.

The main characteristics of probability assessments are listed in Table 4. There are a number of reasons why probability evaluations are often not feasible for assessing programme effectiveness.[10] Firstly, the evaluator must be present at a very early stage of the programme planning cycle to design the randomization. Eligible services, communities or individuals have to be listed and randomized to intervention or control groups. Unfortunately, evaluators are often recruited only well after the programme has been implemented.

It is also necessary to overcome political influences affecting the choice of where to deploy the new intervention. Interventions are usually regarded as desirable and political pressures are put on planners, often resulting in the programme being directed to more influential communities. To ensure the use of random allocation, the evaluator must directly influence the implementation process. Alternatives have been proposed, including the 'stepped wedge design' (or 'experimentally staged introduction'[11]) in which the intervention is deployed in a randomized sequence but eventually extended to all eligible communities or individuals. This eventual extension, as resources become available, is necessary not just for political but also for ethical reasons. This means that randomized designs are not appropriate for looking at effects with long time lags after the intervention begins.

The stringencies of probability trials may result in situations that are artificially different from the reality to which the results must be extrapolated, in other words, that the assessment lacks external validity. The probability assessment may have a high internal validity in showing that the intervention caused the results. But this gain in internal validity may be useless because the lack of external validity renders the results irrelevant to the decisions that need to be made.

Due to those and to other reasons,[10,12] there are many limitations to the use of the probabilistic approach in assessing large-scale programmes. If the intervention has proven efficacy in field trials, few experienced decision makers would require measuring the effectiveness of every programme through a probability design. However, key individuals in donor or international agencies, as well as the evaluators themselves, may have been trained to regard probability assessments as the gold

Table 4 Characteristics of probability evaluations

Type of evaluation	Measurements	In whom?	Compared to what?	Inferences
Probability			Randomized control group(s)	The programme has an effect ($P < 0.05$)
Performance (provision, utilization, coverage)	Programme activities	Implementation workers Programme recipients		Intervention group has better performance than control
Longitudinal-control	Relative change		Comparing before-after between intervention and control	
Impact	Health and behavioural indicators	Programme recipients		Changes in health or behaviour are more beneficial in intervention than control group
Longitudinal-control	Relative change	Target population	Comparing before-after between intervention and control	

standard and fail to understand that this approach is seldom mandatory or even feasible for the routine evaluation of programme effectiveness.

In spite of the limitations of probability assessments there are times when these are essential, such as the first vitamin A suplementation trials that proved the lethality of subclinical avitaminosis A.[13] These early studies had no external validity relative to the implementation of public health interventions even though they were essential to show the need for public health action in populations with subclinical avitaminosis A.

Combining adequacy, plausibility and probability inference objectives

The inference axis has in fact two components that vary together to a large extent. The first component is categorical: adequacy, plausibility and probability evaluations require different designs and result in different inferences, not just in the conclusions to be drawn from statistical tests, but also substantively. The importance of these questions for evaluation design is not discussed in the epidemiological literature.

For instance, a probability influence may deliver a rigorous inference that the intervention caused an impact, without any insight on whether the impact was adequate. Feasibility considerations indicate that some adequacy objectives can be incorporated into the design of plausibility and probability assessments at little added cost. Thus all evaluations should be designed to permit some adequacy inferences.

Logically, there appears to be no advantage of adding plausibility objectives to a probability evaluation. Both are directed at inferring that the intervention had an effect: the first by trying to exclude other explanations for the findings, the second by direct statistical testing. The strength of the inference is greater for probability evaluations, so that there would be not apparent advantage of adding plausibility objectives. However, it turns out that decision makers are not comfortable with a single piece of evidence, no matter how convincing this may be to statisticians or epidemiologists. For example, the exemplary vitamin A probability trials were not believed by many because of lack of congruency.[14–16] This means that some plausibility should be built into probability designs, for example by providing data on confounding variables and, even more importantly, on mediating variables. In the authors' experience, most decision makers are particularly sensitive to evidence of congruency, both from epidemiological data as well as from qualitative components of the evaluation that should complement the former. This congruency is often so persuasive that it may even outweight impact results that do not quite reach statistical significance. It is the congruency of many pieces of evidence that ultimately persuades.

The inference axis has a second component, that is closely related to the first, categorical one. This component reflects the strength of inference about the causality of programme effect. The progression leads from a description without a comparison group, to comparison with possibly biased control groups, and finally to a comparison with a probably unbiased control group (through randomized trials). This second component of the inference axis, unlike the first, is well described in the epidemiological literature[17] and is only briefly discussed in the present paper.

Combining the indicators and the inference axes

Each of the four components of the indicators axis (provision, utilization, coverage, impact) may be assessed according to the three types of inference (adequacy, plausibility, probability). An example is given below in Table 5.

Other factors influencing the choice of evaluation design

In addition to what indicators the decision makers wish to measure and to how certain they want to be, other factors may affect the choice of the appropriate type of evaluation. These include the large-scale efficacy of the intervention, the sector of knowledge to which it pertains, and the timing of the evaluation.

Efficacy

In a perfect world, interventions would only be widely applied at population level after their clinical and public health efficacy had been proven. However, this efficacy is often not demonstrated before practical public health interventions are initiated.

Table 5 Examples of possible evaluations of Diarrhoeal Diseases Control Programmes

Type of evaluation	Provision	Utilization	Coverage	Impact
Adequacy	Changes in availability of ORS in health centres	Changes in numbers of ORS packets distributed in health centres	Measurement of percentage of all diarrhoeal episodes treated with ORT in the population	Measurement of trends in diarrhoeal mortality in intervention area
Plausibility	As above, but comparing intervention with control services	As above, but comparing intervention with control services	Comparison of ORT coverage between intervention and control areas (or dose-response)	Comparison of diarrhoeal mortality trends between intervention and control areas (or dose-response)
Probability	As above, but intervention and control services would have been randomized	As above, but intervention and control services would have been randomized	As above, with previous randomization	As above, with previous randomization

The known efficacy of an intervention, therefore, is another important factor affecting the choice of evaluation design. Let us have two examples of evaluations, from the perspective of international and donor agencies. First, the efficacy of measles immunization is well proven. If adequacy evaluations show that the cold chain is operational and that coverage is high, there is little need for evaluating the impact of immunization programmes on disease rates, or even on changes in immunity to measles. The case is rather different, however, relative to using vegetarian foods to improve vitamin A nutrition. Their efficacy has not yet been established. Demonstration of increased ingestion[18] is insufficient to persuade donors of the utility of this approach without measures of vitamin A status and at least a strong plausibility design. In fact this is a case where more probability designs are likely to be necessary to persuade decision makers to implement these interventions.

Sector of the programme

The subject area of the intervention is another important factor. This paper has concentrated on health and nutrition programmes but the approach can be adapted to other areas. As a general rule, more stringent evaluations seem to be demanded in the health field. For example, health impact evaluations often require the demonstration of a mortality reduction, which will only take place if a number of intermediate changes occur successfully. In other fields, a decision maker may be satisfied with, say, improved performance in a test (in education), an increased crop yield (in agriculture), or greater water consumption (in water/sanitation). In addition, in most other fields the effect is measured solely among the programme recipients, while in health and nutrition more stringent criteria require measurement of coverage or of impact on the whole target population.

Besides differences in the kinds of outcomes measured, distinct sectors require very different degrees of certainty before declaring an intervention as efficacious or effective. Some public policy and programme decisions depend entirely on plausibility statements. This is particularly the case in economics.[19] Even within the health sector there are marked differences in judging the efficacy of interventions, whereby nutritional interventions appear to be held to higher standards than other health interventions.[20]

This variability in standards of certainty required by decision makers in judging the efficacy and effectiveness of interventions is a major barrier to rational public policy. It is therefore important to specify the levels of certainty that are achievable by the designs used, whether they are adequacy, plausibility or probability designs. Comparisons of expected impact for competing interventions across sectors need take these differences in certainty into account.

Timing and timeliness

The time when the evaluation is planned is fundamental. Probability assessments, as noted, require the evaluator to be present before the programme starts so that communities or individuals may be randomly allocated. All longitudinal methods, including those with a control group, also require baseline information to be collected before the programme, or else reliable secondary information for the pre-programmatic period. In general, evaluations of provision and utilization may be carried out sooner and more frequently, as they help local decision makers improve the interventions more quickly than waiting for longer term results. On the other hand, coverage and particularly impact evaluations are often undertaken later in the programme cycle and are often once-off activities. As a general rule, no less than 3–5 years are required for an intervention to show an impact. Several years or decades may be required for showing an impact on diseases with long incubation periods, such as AIDS, chronic diseases or the generational effects of improved nutrition.

As a general rule, evaluations should be planned when the programme itself is being designed, even if actual data collection is only foreseen at a later phase. Adequacy and plausibility evaluations may be instituted after the programme is under way. However, adequacy evaluations are more meaningful if there are clear and feasible pre-set goals, and plausibility evaluations often require baseline information from the pre-programmatic period.

The evaluation should deliver the answers to the decision makers in time for them to take these results into account in their decisions. Perfect information from an ideal evaluation is useless if it arrives after the decision is already made, an all too frequent situation. Therefore evaluators should determine not only what decisions are going to be made but when those decisions will take place. The design and conduct of the evaluation should then be organized to meet these deadlines, and all evaluation designs should include timeliness as part of their objectives.

Magnitude of sampling

The number of people sampled and the distances between them are major determinants of the costs of the evaluation. The number of areas, and the number of people to be sampled within the areas, is determined by calculations based on the willingness of the decision makers to be given erroneous results. Usual practice in scientific research is to accept as true that a treatment has an effect 5% of the time when in fact there is no effect—an alpha error or significance level of 5%. This is an almost sacrosanct figure among academicians. Usual practice in academic research is to declare no effect 20% of the time when there really is an effect—a beta error of 20% usually referred to as a power of 80%. The lower the setting of the per cent alpha and beta errors the greater will be the sample size.

From the above it is obvious that scientists are willing to not identify a beneficial result four times more often than to be mistaken in declaring such a result when it is absent. Most public health practitioners would be very unhappy with that trade-off for evaluating their programmes, and many would set the opposite trade-off. At any rate the sacrosanct 5% significance limit needs to be questioned before being accepted automatically. For instance accepting 20% for both alpha and beta errors would reduce the sample size by 35–40%, below that acceptable to many scientists. Setting explicit per cent error levels that are appropriate for the decision maker is in fact more scientific than blindly accepting conventional levels. Thus one can set the errors much higher in many programme evaluations than in efficacy trials, if the results are not to be used for scientific inferences for which low alpha errors are necessary.

Costs

Costs are often the major factor affecting the choice of a design. Decision makers are particularly sensitive to this aspect, for often they will be asked to provide the necessary funds from the overall programme budget. A full discussion of evaluation costs is beyond the scope of this paper, but evaluators should discuss with decision makers the budgetary implications of different designs, including the following issues:

(a) Is a full summative evaluation worth doing?
(b) Is there a need for collecting new data? If so, at what level?
(c) Does the design include an intervention-control or a before-and-after comparison?
(d) How rare is the event to be measured and how small is the difference to be detected?
(e) How complex will the data analysis be?

Choosing the evaluation design

This section discusses how to combine evaluation designs and also summarizes some of the main points presented above. The classification axes presented above should be used for discussing with decision makers which evaluation design or designs may be used for each programme. Table 6 shows some areas which may typically concern different decision makers in the field of health and nutrition.

Complex evaluations (for example, those with a probability approach or impact assessments) should not be carried out before ensuring, through less costly evaluations, that the process is moving in the expected direction.[3] Table 7 shows a heuristic sequence of evaluations with growing complexity, that would be carried out based on the results of simpler evaluations. It is based on the evaluations listed on Table 5, and contemplates the concerns of local, district and national decision makers.

Table 6 Possible areas of concern of different decision makers

Types of evaluation	Provision	Utilization	Coverage	Impact
Adequacy	Health centre manager International agencies		District health manager International agencies	
Plausibility	International agencies			Donor agencies, scientists
Probability	Donor agencies, scientists			

Note: The **shaded** areas represent those of greater concern for international (e.g. UN) agencies.

Table 7 Hypothetical example of flow of evaluations from simpler to more complex designs, to allow decisions by local, district and national managers

Axes	Provision	Utilization	Coverage	Impact
Adequacy	1st	2nd	3rd	4th (b)
Plausibility			4th (a)	5th
Probability				

One would start by ensuring that ORS is available in the health centres, and next check that the population is utilizing this service. The third stage would include a household survey to assess whether the ORT coverage goal has been reached. So far, all evaluations have been adequacy statements. Next, the decision maker could opt for either showing that coverage is higher in the intervention than in the control areas (option 4(a), a plausibility statement that the higher coverage was due to the programme), or perhaps for attempting to show a reduction in diarrhoeal mortality compared to before the programme (option 4(b), an adequacy statement). This logical sequence helps in deciding the actual sequence, which also depends on funding, administrative and political considerations.

In conclusion this paper is designed to foster the development of a logical framework by which health and nutrition programmes can be judged and compared to other public interventions. The major premise is that the objective of an evaluation is to influence decision makers. How complex and precise the evaluation must be depends on who the decision maker is and on what types of decisions will be taken as a consequence of the findings. Both complex and simple evaluations, however, should be equally rigorous, whether they assess the adequacy of an intervention's effects, or assess the plausibility, or the probability that the intervention caused these effects.

In addition to the above framework, other factors affect the choice of an evaluation design, including the efficacy of the intervention, the field of knowledge, timing and costs. Regarding the latter, decision makers should be aware that evaluation costs increase rapidly with complexity so that often a compromise must be reached.

Acknowledgement

This work was supported by the Evaluation and Research Office, UNICEF.

References

[1] Mohr LB. *Impact Analysis for Programme Evaluation*. Newbury Park: Sage Publications, 1992.

[2] Rossi PH, Freeman HE. *Evaluation*. Newbury Park: Sage Publications, 1993.

[3] Mason JB, Habicht JP. Stages in the evaluation of ongoing programmes. In: Sahn DE, Lockwood R, Scrimshaw NS (eds). *Methods for the Evaluation of the Impact of Food and Nutrition Programmes*. Tokyo: United Nations University, 1984, pp.26–45.

[4] Habicht JP, Mason JB, Tabatabai H. Basic concepts for the design of evaluation during programme implementation. In: Sahn DE, Lockwood R, Scrimshaw NS (eds). *Methods for the Evaluation of the Impact of Food and Nutrition Programmes*. Tokyo: United Nations University, 1984, 1–25.

[5] *New Webster's Dictionary and Thesaurus of the English Language*. Danbury, CT: Lexicon Publications, 1992.

[6] Schlesselman JJ. *Case-Control Studies. Design, Conduct, Analysis*. New York: Oxford University Press, 1982.

[7] Rothman KJ. *Modern Epidemiology*. Boston: Little, Brown and Company, 1986.

[8] Victora CG, Olinto MT, Barros FC, Nobre LC. The recent fall in diarrhoea mortality in Northeastern Brazil. Did ORT play a rôle? *Health Pol Plan* 1996;**11:**132–41.

[9] DaVanzo J, Habicht J-P. Infant mortality decline in Malaysia 1946–1975. The roles of changes in variables and the structure of their relationships. *Demography* 1986;**23:**143–60.

[10] Black N. Why we need observational studies to evaluate the effectiveness of health care. *Br Med J* 1996;**312:**1215–18.

[11] Cook TD, Campbell DT. *Quasi-Experimentation. Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin Co, 1979, pp.375–76.

[12] Kirkwood BR, Cousens SN, Victora CG, Zoysa I. Issues in the design and interpretation of studies to evaluate the impact of community-based interventions. *Trop Med Int Health* 1997;**2:**1022–29.

[13] Sommer A, Tarwotjo I, Dunaedi E *et al*. Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial. *Lancet* 1986;**i:**1169–73

[14] Costello AML. Vitamin A supplementation and child mortality (letter). *Lancet* 1986;**ii:**161.

[15] Gray RH. Vitamin A supplementation and child mortality (letter). *Lancet* 1986;**ii:**162.

[16] Martinez H, Shekar M, Latham M. Vitamin A supplementation and child mortality (letter). *Lancet* 1986;**2:**451.

[17] Rothman KJ, Greenland S. *Modern Epidemiology, 2 edn*. Philadelphia: Lippincott Raven, 1998.

[18] Smitasiri S. On planning and implementing Vitamin A interventions: Linking scientific knowledge to effective action. In: Garza C, Pelletier D (eds). *Beyond Nutritional Recommendations: Implementing Science for Healthier Populations.* Cornell University Press, Ithaca (in press).

[19] Judge GG, Griffeths WF, Hill RC, Lee T-C. *The Theory and Practice of Econometrics.* New York, NY: Wiley, NY 1985.

[20] Hill K (ed.). *Child Health Priorities for the 1990s*. Baltimore: The Johns Hopkins University Institute for International Programmes, School of Hygiene and Public Health, 1992.