AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523
BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY
Batch 65

1. SUBJECT CLASSIFICATION
A. PRIMARY: TEMPORARY
B. SECONDARY

2. TITLE AND SUBTITLE
The application of quasi-experimental methodologies to the design and evaluation of foreign assistance projects

3. AUTHOR(S)
Job,B.L.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1976	104p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS
MUCIA

8. SUPPLEMENTARY NOTES *(Sponsoring Organization, Publishers, Availability)*

9. ABSTRACT
(DEVELOPMENT--DEVELOPMENT ASSISTANCE R&D)

Summarizes investigations made by the Midwest Universities Consortium for International Activities (MUCIA) in 1974 and 1975 concerning the design, evaluation processes and procedures used by A.I.D. in foreign assistance projects. The study attempts to bridge the gap between the practioner and the theoretician concerning quasi-experimental methods by presenting material which more generally treats evaluation and the A.I.D. environment. Methodological discussions are cast within the context of foreign assistance evaluation and design situations. Evaluation, its types and purposes, and its role in the A.I.D. foreign assistance process is discussed, with special emphasis being placed on the importance of sound program and project design. It notes the pitfalls often encountered when social action programs are evaluated. Numerous detailed examples related to situations typical of foreign assistance project and program evaluations are cited. Quasi-experimental and experimental design concepts and methodologies present a series of different types of evaluation designs. These designs explain how each avoids threats to validity and is adaptable to use in foreign assistance projects. The report summarizes arguments for the adoption by A.I.D. of more rigorous design and evaluation methods. Actual project designs demonstrate how quasi-experimental and experimental principles may be combined in a single design format.

10. CONTROL NUMBER
PN-AAE-051

11. PRICE OF DOCUMENT

12. DESCRIPTORS

13. PROJECT NUMBER

14. CONTRACT NUMBER
CSD-2958 211(d)

15. TYPE OF DOCUMENT

AID 590-1 (4-74)

211(d)
MUCIA PN-AAE-051

THE APPLICATION OF QUASI-EXPERIMENTAL METHODOLOGIES TO THE DESIGN AND EVALUATION OF FOREIGN ASSISTANCE PROJECTS

BRIAN L. JOB

Department of Political Science
University of Minnesota

November, 1976

Note: This is a revised version of a monograph originally prepared in 1975 in conjunction with a research grant received from MUCIA, with Frank Hoole of Indiana University, to study foreign assistance project design and evaluation

Introduction

This monograph summarizes investigations undertaken by the author concerning the design and evaluation processes and procedures utilized by the U.S. Agency for International Development (AID). Specifically, this research considers the possibilities for the application of quasi-experimental design, and to a lesser extent experimental design, techniques and methodologies to the design and evaluation of foreign assistance projects.

Funding for this effort was provided by the Midwest Universities Consortium for International Activities from a larger grant from AID (MUCIA/AID 211(d) Grant, AID/csd-2958-14). Work was commenced during the summer of 1974 in conjunction with Professor Frank Hoole of Indiana University, a first draft of the monograph was completed during the summer of 1975, revisions were made during the summer and fall of 1976. Material for this report was gathered from the extensive literature on evaluation and research design of domestic social action programs; from AID manuals, reports, and memorandum; and from interviews with AID officials. Those portions of the monograph detailing the various different types of quasi-experimental designs present, in an expanded fashion, material which I utilized in a series of lectures on the same subject to the AID Development Officer Training Program. I would like to particularly acknowledge the assistance of Herbert Turner, of the Agency of International Development, Richard Blue, of the Development Officer Training Program, AID; and Grafton Trout, International Development Research Center, Indiana University. These individuals, of course, are not responsible for any errors or omissions in my work and do not necessarily share my opinions on some of the issues discussed. I am indebted, also, to Barbara Freeman who with great skill and much patience typed the various drafts of this monograph.

A few words are in order about the purpose and format of the monograph. I have deliberately taken a broad perspective toward the task at hand and consequently have included considerable material which does not specifically present quasi-experimental methods but which does more generally treat evaluation and the AID environment. The statistical and methodological intricacies of experimental and quasi-experimental are well-developed and are presented in numerous academic texts and journal articles. However, without interpretation such technical and theoretical statements mean very little to the governmental decision-maker or field officer. I have attempted, therefore, to bridge the gap between the practicioner and the theoretician by casting the methodological discussions within the context of foreign assistance evaluation and design situations. This has not been an easy task; finding or devising examples that are relevant without being too complex, or that are useful without being simplistic proved to be challenging and thought-provoking.

Just as I do not see this monograph as a substitute for a methodology or a design text, neither do I see the monograph as a replacement for the various governmental publications on the topic of evaluation--specifically the Evaluation Handbook and the Project Evaluation Guidelines produced by AID. It is instead to be supplement to these publications, providing additional information about quasi-experimental and experimental design procedures.

There are four chapters in the monograph. A short introduction is given in the first of these to the concept of evaluation, the different types and purposes of evaluation, and the role of evaluation in the AID foreign assistance process. The crucial nature of sound program and project design as a prerequisite to successful and useful evaluation is emphasized. Chapter Two proceeds in a somewhat different fashion by discussing the

various pitfalls that can be encountered when evaluating any social action program. Numerous threats to validity, as they are called, are explored in some detail along with examples related to situations typical of foreign assistance project and program evaluations. This paves the way for the detailed exposition in Chapter Three of quasi-experimental and experimental design concepts and methodologies. A series of different types of evaluation designs are presented with explanations as to how each avoids or is susceptible to certain threats to validity and how each is more or less adaptible to utilization in foreign assistance project situations. Chapter Four summarizes the major arguments for the adoption of more rigorous design and evaluation methods by AID and other social program agencies. Several actual project designs are outlined to further demonstrate how quasi-experimental and experimental principles may be combined in a single design format.

Chapter I
The Evaluation Process

A. Evaluation: Definitions and Process

United States governmental sponsorship of, or participation in, national and international programs designed to alleviate undersireable conditions or to raise standards of health, education, welfare, or prosperity among selected domestic or foreign populations has increased dramatically during recent decades. With the proliferation of these so-called social action programs has naturally come a concern on the part of legislators, administrators, and participants as to whether such programs are actually producing results, whether or not these results are significant, and whether or not the expenditures of time and effort involved can be made more efficiently. The task of providing answers to such questions is termed evaluation. Evaluation activities have come to compromise a major component of the decision-making process in all federal agencies. This study focuses specifically on the evaluation of programs and projects under the direction of the Agency for International Development.

The term "evaluation" may be generally defined as any manner of assessment of a phenomena by appraisal or comparison with some standard. In the context of this study, a more restrictive definition is required--one that is specifically oriented toward bureaucratic decision-making processes. The Office of Management and Budget, for example, has defined evaluation as "an activity which produces and subjects to systematic impartial analysis, information concerning the impact, effectiveness, or efficiency of Federal programs in achieving their objectives."[1] In a similar fashion, the AID Evaluation Handbook describes evaluation as "a systematic assessment of

actions in order to improve planning or implementation of current and future activities."[2] It is important to take note of two essential characteristics of each of these definitions.

The first concerns the purpose of evaluation. Evaluation is intended for use; it is a utilitarian activity, designed to provide decision makers with up-to-date, accurate information so that wise decisions may be taken to strengthen and modify existing programs or to design and implement new programs.[3] The success of evaluation activities themselves, therefore, is directly related to the extent to which the decision-maker is provided with answers to his questions. Evaluation is a part of the ongoing management process within any agency that also entails planning, implementation, and administration of programs and projects. As such, evaluation also becomes a political activity: the viewpoints, questions, and demands of decision-makers, particularly those at different positions and levels within the organization, may differ and thus subject the evaluator to cross-cutting pressures and influences. The political and bureaucratic aspects of his task, however, should not be allowed to detract the evaluator from his mission of proceeding in a scientific fashion to produce as complete, accurate, independent an assessment as is possible.

The second important feature of evaluation concerns the nature of the activities involved in such an enterprize. Edward Suchman, for instance, in characterizing evaluation as "measurement of effectiveness in reaching some predetermined goal" identifies three key elements in his definition.[4] These are (1) objectives or goals which decision-makers consider desireable and of positive value, (2) a planned program of intervention which the decision-makers believe is capable of achieving the desired goals, and

(3) a method for determining the degree to which the desired objectives are attained as a result of the program.

It is the implementation of this third element, i.e. a method for determining..., which challenges the evaluator. In order to construct his assessment he must (a) identify the goals of the decision-makers and design operational indices of them, (b) analyze the logic of the project design and the manner in which the project was put into operation, (c) devise operational indicators of the outcomes of the project, (d) measure the extent of change which has taken place because of the project, (e) establish that this change may be attributed solely to the effects of the project and finally (f) make projections and suggestions as to the operation of the project in the future. The natural logic of evaluation is, therefore, comparative in nature. Indicators of conditions before the project are compared to indicators of conditions after the project to discover if the project worked, i.e. a matter of validity. These indicators are then compared to indicators of the decision-makers goals to discover if the results are "good" or useful, i.e. a matter of value.

B. Evaluations: Types, Purposes, and Criteria

An evaluation, if it is to fully satisfy its function, must ask and attempt to answer the following questions:[5]

1. What is the nature of the objectives? That is, what is it that the decision-maker wants the program to accomplish? Is there a single objective or several?

2. What is the desired magnitude of changes to be achieved, i.e. what is the definition of success or significance?

3. How are the objectives to be achieved? What is the logic, design, and implementation of the project?

4. Who is the target of the program? Is it individuals, groups, villages, communities, for example, that are to be affected?

5. When is the desired change to take place? Are the effects of the program to be short-term, intermediate, or long-term? Are there side-effects?

6. What are the observable results of the project? How are they measured? Are they "significant"?

7. How do these outcomes and assessments affect future decision-making concerning the project in question? new projects?

This listing, of course, makes it clear that evaluation is not a simple task. Answers to these questions may be extremely hard to find: empirical indicators may not be well-chosen or sensitive to project outputs, effects may be so long-term or short-term in nature so as to escape detection, project inputs may not be delivered accurately or completely, etc. Furthermore, there may be no agreement among the decision-makers themselves as to the nature of the goals, or the out-puts of the project, or the degree to which the outputs satisfy the attainment of the goals.

Evaluation of a single program or project may be undertaken for a variety of reasons, by decision-makers at differing levels within the organization utilizing different techniques or measures. In every instance, the results and the interpretation of an evaluation of the program or project may be quite unique. A more complete understanding of the various types of evaluation, reasons for evaluation, and levels of decision-makers who commission or perform evaluation studies in a large bureaucracy such as a federal agency will help the reader to appreciate the complexity of evaluation design.

One sheme by which the differing types, levels, and purposes conducted by the federal government may be categorized has been devised by Joseph Wholey of the Urban Institute.[6] A brief general review of his ideas will as a good introduction to our more specific treatment of evaluation as performed by the Agency for International Development.

Wholey makes an initial distinction between the programs and projects of the federal government.

> A federal program is defined as the provision of federal funds and administrative direction to accomplish a prescribed set of objectives through the conduct of specified activities.
>
> A project is the implementation level of a program--the level where resources are used to produce an end product that directly contributes to the objectives of the program.[7]

A program, therefore, is designed with broad objectives and large amounts of funds and resources that are to be apportioned to projects designed to further the overall programatic goals by implementing specific objectives in selected locales. Thus, the national Head Start program administered by the federal government encompassed numerous Head Start projects in cities and local school districts. A somewhat similar distinction exists in AID as well: a program to increase the productivity of the small farmer in a certain nation may involve numerous projects, some dealing with irrigation, some with introducing new varietals, some with extension services, and some with the establishment of rural credit unions. The "successful" implementation of each project would presumably bring the overall program closer to realization of its objectives.

Wholey then identifies four major types of evaluations:[8]

1. program impact evaluation: assessment of the overall effectiveness of a national program in meeting its objectives, or assessment of the relative effectiveness of two or more programs in meeting common objectives.

2. program strategy evaluation: assessment of the relative effectiveness of different techniques used in a national program.

3. project evaluation: assessment of the effectiveness of an individual project in achieving its stated objectives.

4. project rating: assessment of the relative effectiveness of different local projects in achieving program objectives.

Evaluations of programs, i.e. 1 and 2 above, will naturally involve different personnel and different techniques, and different criteria of success than will evaluation of projects, i.e. 3 and 4 above.

Consider for a moment the following five categories of criteria according to which the success or failure of a program or project may be evaluated.

a. effort: the quantity and quality of activity that takes place, an assessment of inputs without regard to output.

b. effectiveness: the assessment of the results of the effort rather than the effort itself; a comparison of the results to the objectives of the project or program.

c. impact: the degree to which success is adequate to the total amount of demand or need.

d. cost effectiveness: a comparison of alternative methods in terms of cost; on assessment of the ratio between effort and resultant impact.

e. process: an assessment of the positive and negative side-effects of the project or program--both anticipated and unanticipated. "This is not an inherent part of evaluative research but rather an analysis of the process whereby a program produces the result it does...."[9]

Evaluations of different types will involve different levels of decision makers who will give higher or lower priority to the various evaluation criteria. Thus, program impact evaluation will involve high level agency officials and program directors who will be concerned with effectiveness, impact, and cost effectiveness. Program strategy evaluations will similarly involve high level personnel but their interests will now be explicitly comparative in nature. Therefore, effectiveness, cost effectiveness, and process will be major considerations. Lower-level decision-makers, that is program managers, project directors, and project personnel will participate in project rating and project evaluations. The important criteria in these instances will be effort, effectiveness, and cost effectiveness.

With a bit of oversimplification, the key distinctions suggested by Wholey's scheme may be summarized in the Figure I below:

Figure I

criteria of evaluation	type of evaluation	level of administrator
effort	project evaluation	project directors program managers
effectiveness	project rating program strategy program impact	program managers
impact	program impact	program directors agency decision-makers
cost effectiveness	program strategy program impact	program directors agency decision-makers
process	program strategy	program managers

While the exact details of this categorization may not be critical, what is of major import is the understanding that evaluation is not a uni-dimensional phenomenon. There is no pat formula or design which can be implemented in every instance. If the decision-makers' questions concern whether or not television is an effective educational tool, for example, a quite different evaluation design will be required than if agency officials want to know if the Head Start program had significant impact.

Agreement must also be reached on matters of goals and objectives if evaluation is to be undertaken. An assessment of cost-effectiveness or impact will only be possible if the evaluators and decision-makers have a

common understanding of the standards or bench mark against which comparisons are to be drawn. To the program manager in a international agricultural program, for instance, the fact that small farmers have adopted new techniques and varieties may be far more significant than the fact that productivity increased slightly. However, to the program director and to the agency decision-maker who must deal with Congress, large increases in productivity are a prime objective since these are straightforward and visible indications of agency success. Establishing a basis of understanding on goals and objectives, therefore, becomes one of the major tasks of an evaluator.[10]

It should be noted that activities of a project monitoring nature are not included as a separate type of evaluation in this discussion. While monitoring, auditing, and reporting of project or program delivery is a necessary aspect of implementation, these tasks are not usually termed evaluative. This is because they relate only to measures of input and have no necessary relationship with measures of output.[11] Evaluation, on the other hand, assumes delivery of at least some inputs and is based on a comparison of these to the outputs of the project or program. The AID Evaluation Handbook expresses this view-point quite strongly in stating that,

> evaluation...differs materially from monitoring or from regular audits and inspections. The latter are generally designed to appraise operations in order to determine compliance with management controls and regulations. As such, they do not as a rule challenge the choice of targets.[12]

A good monitoring and auditing system will certainly make the task of the evaluator much easier by providing him with information about the schedule, cost, and delivery of inputs and the production of outputs. However, judgements as to the effectiveness and efficiency of the project or program involve careful design, measurement, and testing and thus are part of evaluation, not monitoring exercises.

C. Evaluation and the Agency for International Development

The Agency for International Development, unlike most federal agencies, conducts many of its activities in foreign countries. The scope, size, and cost of these operations varies enormously. Evaluation of programs and projects has always been an integral part of AID management. However, with Congress, AID directors, and the public giving greater critical attention to governmental expenditures on social action programs--especially those in foreign countries--the design and conduct of evaluations is becoming increasingly more important. Before proceeding with our discussion, a brief summary of the different types and levels of evaluation conducted within AID should prove useful. For a detailed treatment of the agency's policies and procedures, the interested reader should consult the AID Evaluation Handbook, and AID Project Evaluation Guidelines.[13]

The distinction between programs and projects noted earlier exists within AID operations as well, with the qualification that "program" is used to refer to all activities within a particular country. A program therefore, will encompass the activities of the numerous projects sponsored by AID in one nation. Evaluations undertaken by the agency occur at four levels: project evaluations, sector level evaluations, program level evaluations, and country level evaluations.[14]

1. project evaluations: There are three kinds of AID projects: technical assistance projects, capital assistance projects and food assistance projects. Capital assistance projects, such as the granting of loans, are planned and assessed according to quite specific economically-oriented criteria. They will not be discussed in this study. Non-capital assistance projects (health, agriculture, welfare, education projects, for instance) are supposed to be routinely evaluated each year. In special cases, in-depth evaluation studies are commissioned for project assessment.

2. sector-level evaluations: Evaluations of this nature attempt to survey the needs of, or discover the impact of, AID programs and projects upon a chosen sector within a foreign country. For example, a sector-level evaluation may be commissioned to evaluate the impact upon the agricultural sector of all AID activities in Columbia during the last five years.

3. program-level evaluations: "A country program evaluation consists of reviewing the significance and success of all AID developmental activities within a particular country." These evaluations which occur infrequently are undertaken when a comprehensive review is required with an eye towards replanning strategy or levels of assistance.

4. multi-country evaluations: Evaluations of this type are comparative in nature. They attempt to examine development processes and the effect of foreign assistance upon these processes in a number of countries. Examples of such studies would include an evaluation of AID's use of program loans to influence the economic policies of developing countries, or an analysis of extension services in Latin America.

It is apparent that AID categorizes its evaluations as to level and subject matter rather than as to purpose--in contrast to the Urban Institute scheme described by Wholey. An AID project evaluation, for instance, could be designed to assess either effort, i.e. an evaluation of inputs, or effectiveness, i.e. an evaluation of results, or cost effectiveness, i.e. an evaluation of different methods or techniques. Project personnel in the field and officials of the host country would most often conduct studies of this type.

An AID sector-level evaluation, or an AID program-level evaluation, on the other hand, could be commissioned to assess effectiyeness, impact, cost effectiveness, and/or process depending on the questions being asked and the type of program or projects involved. For example, the agricultural sector-level evaluation of Columbia would probably focus upon the criterion of impact, i.e. the relationship of results to need, while an evaluation of family planning activities in Guatemala may focus on cost

effectiveness to try and determine which combination of instruction and technique were most successful. Almost certainly, higher level agency personnel and host country officials would be involved in evaluation of this nature.

Multi-country evaluations constitute quite extensive undertakings, as for example the AID land reform review that examined thirty countries, of which about one half were AID aid recipients.[15] Again the results of such evaluations are going to interest decision-makers at the highest administrative and policy making levels. Criteria of impact, process, and cost effectiveness will be important. Decision-makers, to continue the example, will want to know in which countries AID programs had the most impact, which types of projects worked best, etc.

This overview of AID evaluation serves to emphasize the range and complexity of the evaluation activities within the agency. Three major factors will tend to ensure this continued complexity. First, the initiation, design, and implementation of projects and programs requires the cooperation and coordination of decision makers at many levels within the agency and within the host country bureaucracy. Second, AID activities are pursued in relatively uncertain environments. Third, interpretation of the goals and benefits of a project or program may vary considerably among AID Washington decision makers, AID field personnel, host country officials, and host country participants. A "successful" project for one group may be far less successful to another.

Systematic, scientific evaluation of AID activities is a prerequisite to the continued success and usefulness of the agency's programs and projects. Increased attention must be given to increasing efficiency, effectiveness, and impact of program efforts and evaluation helps to provide such information.

It is clear, however, that evaluation is a challenging task in such an environment. Projects and programs can not be evaluated without the expenditure of considerable resources and careful planning and design. Thus, the place of evaluation in the larger management process of planning, design, and implementation will be considered next.

D. Evaluation and Design

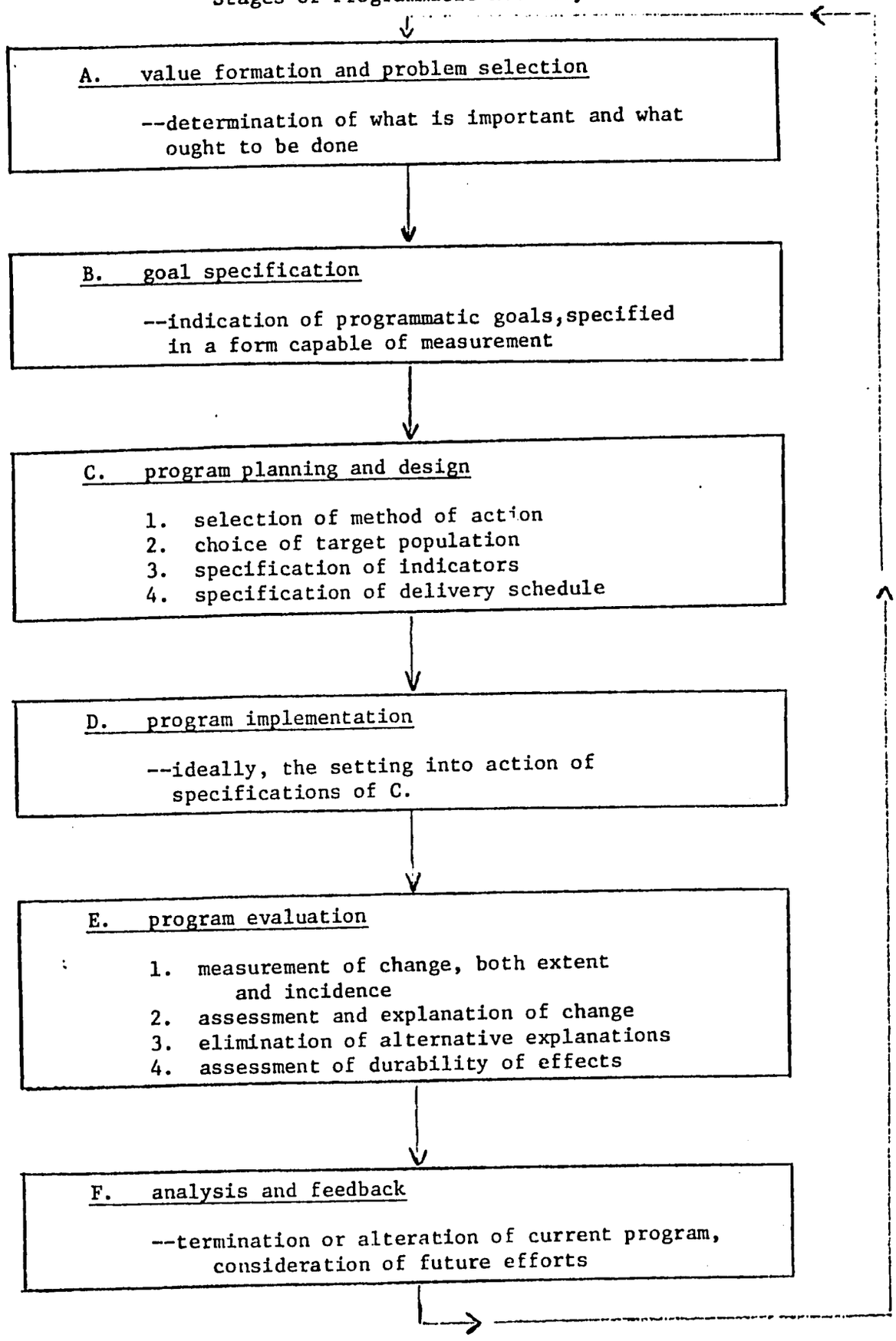
To the administrator, program evaluation is but one of the activities involved in the progression from the time that the program is only an idea until the time that the program is officially ended. One could view this process as entailing at least six different logical steps as seen in Figure II.[16]

Figure II about here

AID's planning, evaluation, and design of programs and projects follows a pattern quite similar to that outlined in Figure II. Those faimilar with so-called "log-frame," for instance, will detect the correspondence between the tasks noted in the Figure and those specified in the log-frame schema. (The logical-framework is the name given to a planning document which must be filed prior to approval of any new agency project.)[17] The requirements that goals, purposes, and targets be made explicit, that indicators be formulated in measureable terms, that budgets and schedules be specified, and that assumptions and causal linkages be stated are all contained in the log-frame. AID decision-makers realize that the log-frame is not an evaluative device in itself; it is instead a research design tool. However, these decision-makers also realize that evaluation is dependent on research design. For by indicating clearly what a project wishes to

Figure II

Stages of Programmatic Activity



accomplish, how results are to be calculated, and how certain conditions in the environment may affect project implementation in advance, the log-frame provides the evaluator with a valuable starting point.

While the Figure presents an admittedly abbreviated picture of complex decision making and administration, it does serve to highlight several crucial points concerning evaluation. First of all, evaluation should not be viewed as a post-planning, post-program activity--something that is contemplated and performed after the program activity itself is underway. Evaluation, if it is to be useful and productive, must be planned in advance. Once subjects have been chosen, personnel are in the field, and data have been collected, it is often too late to realize that there are no measureable goals to the program activity, or that the wrong group of farmers were chosen, or that the data do not really reflect the right indicators. These matters should be thought out in advance. Whenever possible disagreements as to goals should be thrashed out prior to implementation and evaluation. If adequate attention is given to devising indicators of goals, of conditions before the program, and of conditions during and after the program then evaluation becomes a much easier task.

Completion of a successful evaluation, i.e. one that produces accurate and relevant information, is thus highly dependent upon execution of the prior steps in the programmatic process. This position receives strong support from Bernstein and Freeman, who, upon surveying federal evaluation studies, found most of them to be unscientific and unsystematic in nature. In their view, the essential flaw in almost all of these programs could be traced to inadequate research planning and design.[18]

The second important point illustrated by the diagram is that evaluation does not mark the end of the programmatic process. Evaluative studies are really only worth doing when the knowledge which is gained from them is

fed back into the decision making process. If a project or program is assessed at its conclusion, information should be gathered to aid in the planning of future undertakings. If evaluation is done while activities are underway, e.g. an annual review of a five-year AID project, weaknesses may be spotted that may be corrected immediately. To reiterate an earlier point, the purpose of evaluation is to aid decision-making. Unless evaluation is performed at times and in a manner which provides useful and accurate information to decision makers, it is of limited utility.

E. The Feasibility of Evaluation

As a conclusion to this section of the study, it may prove useful to briefly consider under which circumstances evaluation will be most feasible and most likely to produce useful results. Weiss, Boruch and Reicken, and Wholey, who have all devoted some thought to this question, regard the following factors as conducive to satisfactory performance of evaluation:[19]

1. There exists some doubt in the mind of the decision-maker as to the efficiency, effectiveness, or impact of the program in question. That is, the decision-maker is interested in the results of an evaluation.

2. There is reasonable assurance that the program, or subsequent programs, will actually be affected by the results of the evaluation. In other words, the program is not, for political, social, or economic reasons, perceived to be inalterable.

3. There exists methods and resources which will allow collection of data, determination of goals and indicators, and comparisons necessary in an evaluation. (Note that the subsequent sections of this study focus upon the types of evaluations which may undertaken in varying circumstances. The statement of this third point is therefore to emphasize that some minimum of data and information is prerequisite to evaluation.)

4. There are adequate managerial, technical, and field staff along with adequate financial resources available with which to conduct an evaluation.

NOTES

[1]"Federal Program Evaluation from the OMB Perspective," Public Administration Review, vol. 34, no. 4, July/August, 1974.

[2]Evaluation Handbook, 2nd ed., Office of Program Evaluation, USAID, Washington, May, 1974, p. 2.

[3]Carol H. Weiss, "Alternative Models for Program Evaluation," in William Sze and June Hopps, eds., Evaluation and Accountability in Human Service Programs, Cambridge, Mass.: Schenkman Publishing Co., Inc., pp. 113-126.

[4]Edward A. Suchman, "Action for What? A Critique of Evaluative Research," in Carol B. Weiss, ed., Evaluating Social Action Programs: Readings in Social Action and Education, Boston: Allyn and Bacon, Inc., p. 53.

[5]This list represents an adaptation and expansion of the list of questions given by David Banner, et al, in their The Politics of Social Program Evaluation, Cambridge, Mass.: Ballinger Publishing Co., pp. 15-16.

[6]Joseph S. Wholey, et al, Federal Evaluation Policy: Analyzing the Effects of Public Programs, The Urban Institute, Washington, D.C., 1971.

[7]Ibid., p. 24.

[8]Ibid., pp. 24-26.

[9]This list of criteria, which was originally formulated by Edward Suchman, is taken from Wholey's monograph, ibid., p. 94.

[10]For a further discussion of this point see Edward Suchman, op.cit., p. 67.

[11]Joseph Wholey, et al, op.cit., p. 95.

[12]Evaluation Handbook, op.cit., p. 3.

[13]Ibid., and Project Evaluation Guidelines, 3rd ed., Office of Department Program Review and Evaluation, USAID, Washington, D.C., August, 1974.

[14]Evaluation Handbook, op.cit., pp. 21-30.

[15]Cited in ibid., p. 30.

[16]The steps outlined in Figure II represent an expansion of Suchman's discussion of the evaluation process, as it is presented in David Banner, et al, pp. 14-15.

[17]Refer to Project Evaluation Guidelines, op.cit.

[18]See I. Bernstein and H.E. Freeman, Academic and Extrepreneurial Research: The Consequences of Diversity in Federal Evaluation Studies, New York: Russell Sage, 1975.

[19]Carol B. Weiss, Evaluation Research, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972, pp. 10-11; Robert Boruch and Henry Riecken, "Applications of Randomized Experiments to Planning and Evaluating AID Programs," Final Report of AID Grant no. AID/cm/ta-c-1055, October, 1974, pp. 30-33; and Joseph Wholey, et al, op.cit., pp. 77-85.

Chapter II
Evaluation and Threats to Validity

A. Three Aspects to the Evaluator's Task

Attention for the moment will be confined to project level evaluations because most social science scholarship on the topic is directed at this level and also because this level constitutes the lowest rung in AID's evaluation scheme. A clear understanding of evaluation at the project level will provide a stepping stone to consideration of evaluation at the program, sector, and country level.

The following example may be used to illustrate several terms and phrases which occur throughout the evaluation literature: Tropicana is a small, poor, largely agricultural nation in Central America that maintains good realtionships with the United States. The Agency for International Development in an attempt generally to improve economic conditions in Tropicana and more specifically to increase small farmer productivity has set up a project within one province to establish a series of rural cooperative credit unions. Assume that the project has run for some four or five years, and that AID decision-makers wish to evaluate the project to assess its effectiveness and impact. This hypothetical situation would therefore be typical of many AID operations abroad.

Adopting evaluation terminology, the project could be characterized as involving inputs, outputs, goals, a target population, and an assumed causal relationship. The inputs would be the trained personnel, the initial capital for the credit unions, the materials to build facilities, extension officers, etc. That is anything which the agency might

provide to the project, could be termed an input. Note that the set up and delivery of this package of inputs would be called an intervention (or in psychological terms, a treatment). AID, in other words, is "intervening" in a deliberate way in the environment of the farmers in an area of Tropicana. The particular province chosen, and the specific villages and farmers within these village communities for whom the project is designed would be called the target population. Outputs of the projects may include such items as number of credit unions put into operation, number of loans granted and/or paid back, increased average annual yield per farmer in the target population, and increased average annual incomes of these farmers. The goals of the project, when stated in measureable terms, might be a twenty per cent increase in grain yields over a five year period, the establishment of fifteen active, operational credit unions, a rise of ten per cent of small farm family income, etc. Underlying the design of the project is the assumed causal relationship that provision of capital to small farmers will provide them with needed resources to obtain new seed, equipment, or land and that consequently their productivity and incomes will increase. Each of these factors will play a part in any evaluation of this project.

The task of the evaluator is to determine to what extent the results of the project satisfy certain criteria set out by the decision makers. (Recall the five criteria of success discussed earlier.) Thus if the decision maker is interested in assessing the "effectiveness" of the Tropicana project, he is interested in comparisons of the observed (presumably positive) increase in production to the project goal of a certain specified percentage

increase. In other words, the evaluator seeks to determine if the assumed causal relationship between implementing the intervention program and attainment of the desired goals was true for the project.

There are actually three aspects to the evaluator's job. First, the evaluator must establish by empirical verification, i.e. the collection of information and data, that the inputs actually were delivered. In the example, this would entail establishing the credit unions were set up, determining how many loans were made and to whom, gathering crop information records, etc.

Second, the evaluator must demonstrate the existence and the nature of the cause-effect relationship exhibited in the project. This really involves three sub-tasks: (a) the determination of an observed change in the level or rate of the project outputs, e.g. yields per acre on small farms rose five per cent; (b) the determination that the amount and direction of the observed change was meaningful and significant, i.e. the results could not reasonably be dismissed as a chance occurrence; and (c) the demonstration that there is a relationship between the effects of the project inputs and these observed changes in the outputs.[1] For instance, the evaluator would like to show that in regions where no credit unions were set up, productivity showed no increases and/or in those regions where credit unions were established, overall productivity increases varied according to the volume of credit union loan activity.

Third and finally, after isolating an input-output relationship in the project and assessing the extent and magnitude of this relationship, the evaluator must rule out all other conceivable factors which might have brought about the same effects. That is, he must systematically consider

and eliminate all rival hypotheses which could account for the observations but which were the result of factors outside the control of the project personnel. In the literature on evaluation, these confounding factors are referred to as threats to validity.

In the credit union project example, rival hypotheses which might explain an increase in production could include such situations as: a marked improvement in the weather over previous years, an inaccurate or dishonest system of measuring agricultural output, or the fact that the loans were actually administered only to larger farmers who were already efficient producers. To the casual observer, the project in each of these circumstances might appear to be successful. However, such a conclusion would be invalid, and the decision-maker who attempted to implement similar projects elsewhere, or to increase the scope of the present project hoping to achieve even more marked results, would be inviting major setbacks. Of the three aspects to the evaluator's job, the most taxing is the systematic investigation and elimination of rival hypotheses. Finding evidence of change and establishing a plausible connection between the inputs and outputs will usually not be too difficult in a technical assistance project. Establishing beyond a reasonable doubt that it is indeed the intervention of the project which accounts for the changes will be a challenge to any evaluator.

Many threats to validity can be forestalled by careful planning in the initial design of the project or program. Design is in many ways half the battle in evaluation. Unless the project is set up in advance to discount the effects of certain confounding influences, no amount of ex post facto analysis can remove them. One way to appraise different types of research design is to catalog all the possible threats to validity and

discuss each research design in the context of its susceptibility or resistance to those threats. As can be anticipated, there is no perfect research design and no foolproof evaluation procedure. While some designs are obviously better than others, each different type of design will be subject to at least some of the threats mentioned below. If the evaluator is consulted in advance, he may suggest ways of eliminating future problems. Even if circumstances dictate the adoption of a less than optimal project design, an evaluator, by being aware of the possible dangers, can temper his findings accordingly.

B. Questions of Validity in Evaluation

The confirmation of a causal relationship within a project or program, the elimination of all rival hypotheses, or confounding factors as they are sometimes termed, which might also account for an observed relationship are questions of validity-- "validity" in the sense that the evaluator or decision maker draws accurate and true conclusions about the nature of the events and interaction which he has observed or managed. The matter of validity is thus of prime concern to the evaluator and a topic much discussed in the literature on research design and evaluation. Most prominent, of course, are the writings of Donald Campbell and his associates on this subject.[2]

When considering the implementation of social action programs, Cook and Campbell, discuss four major types of validity to be considered in design and evaluation.[3] These are internal validity, external validity, construct validity, and statistical validity. Of the four, perhaps the

most important is the first, internal validity. Internal validity obtains when correct conclusions are drawn as to the nature of the causal relationship between the inputs, A, and the outputs, B, in the context of a specific project or program. Alternately, internal validity does not exist, or is threatened, when the evaluator attributes a causal relationship between A and B that is actually the result of another factor C, operating within the context of the project or program, which is producing the observed effect upon B. The prescence or the operation of the factor(s), C, is either unknown to, or uncontrolled by, the project manager. When investigating internal validity, the evaluation thus has to examine the credibility of all alternative explanations which could have arisen while the project or program was underway. Internal validity is distinguished from the three other types of validity in that it deals with the elimination of possible interferences or factors operating in the environment within which the project was designed and implemented.

External validity, on the other hand, concerns the extent to which the evaluator may generalize to other settings, populations, and times on the basis of the single project that he has analyzed. Thus while internal validity relates to the environment in which the project was implemented, external validity relates to the other environments in which the project might be implemented. This is then a matter of generalizability. Any aspect of the project environment, target population, or implementation which was unique to that particular situation, therefore, would be a threat to external validity in that these circumstances may not be replicated elsewhere. Setting up the Tropicana project, for example, in a province populated by farmers whose religion, race, or ethnicity differed substantially from those in other areas of the country would severely restrict the external validity of the project. The decision maker would have little

assurance that a similar project would succeed if set up in an area populated by a quite different class of farmers.

Construct validity relates to whether or not the assumed causal relationship underlying the project is conceptually sound. In terms of the Tropicana credit union example, construct validity is at issue when someone questions the assumption that increased availability and use of credit by the small farmer will result in his becoming a more efficient producer. Threats to internal validity are different from threats to construct validity in that the former are factors that disturb the causal relationship in the project as implemented. The latter, however, lead to misconceptions because the project was incorrectly planned and designed prior to implementation. Cook and Campbell nicely distinguish between external and construct validity in the following statement:

> Those aspects that focus on generalizing across times, settings, and persons are retained under external validity and those that have to do with matching the operationalized treatments and measures to theoretical constructs (that is, those that focus on generalizing across treatments and measures) are grouped under construct validity.[4]

In foreign assistance projects, attention has often given to more applied considerations; the "larger purpose" of the projects is rarely questioned. In reference to AID, however, this bonus will likely change because the recent congressional directive that foreign assistance must demonstrably improve the lot of the rural poor will force a major re-examination of the conceptual frameworks, i.e. the construct validity, of many previous agency activities.

The fourth and final class of threats involving statistical conclusion validity concerns the appropriate use of statistics and statistical tests in evaluation. Investigators use statistical techniques for two reasons when analyzing project data: (a) to detect if changes in the outputs of the treated group are significant and (b) to detect if the differences, if any, between the outputs of the target population of the project and the outputs of populations outside the project are significant. In our example, (a) would refer to an assessment of the increases of production of farmers in the credit union project, (b) would refer to a comparison of these increases to the production figures of other farmers in Tropicana.

In statistics, the word significant refers to the degree to which observed differences depart from differences which could have been expected strictly on the basis of chance. A level of significance, usually around 1% or 5%, is set by the investigator, meaning that he will accept as "true" any result which statistical tests indicate could only have occurred by chance 1 or 5 times out of 100. Threats to statistical conclusion validity relate to circumstances which might lead the evaluator to accept results as "true" which are not, or to over or underestimate the statistical significance of the observed outputs. Because threats to statistical validity require quite technical explanations, a list of them will not be given below in this study.[5]

The decision maker should most certainly have his project data analyzed statistically. Competent analyses can provide evidence about the direction and degree of changes in the target population which might go undetected by a causal observer. An evaluation team, therefore, should contain at least one person who is technically skilled in the statistical analysis of data. Furthermore, other members of the team should have an understanding of the principles, if not the details, of the statistical

tests and procedures which are used. Sophisticated statistical analysis can occasionally provide right answers to wrong or irrelevant questions. The problem can be alleviated if the person in charge of the evaluation formulates his requirements precisely with a rough understanding of what statistics can and can not do for him.

While seemingly a negative way to introduce a discussion of research design, we now turn to a list of specific threats to internal, external, and construct validity which might occur in typical foreign assistance project evaluations. This catalogue will serve to illustrate the distinctions drawn above and it will provide the reader with a better understanding of the pitfalls to be avoided when planning and designing projects or programs. To someone with experience in evaluation of social action programs many of these so-called threats to validity will have been already encountered, although probably not with the particular labels and categorization employed below.

In the lists below, the reader will find references to projects with a control group(s). This term will be discussed more fully in the next section of the monograph dealing explicitly with the design of evaluations. For the moment, a control group can be viewed as a separate group of individuals who do not receive any inputs (or who receive inputs different from those received by the target population). Control groups should be composed of individuals, villages, farmers, etc. similar to those in the target population so that any differences observed in the results or behaviors between control and target groups may be attributed to the effect of the project. In the Tropicana example, a control group may consist of a group of farmers in a neighboring province whose production figures are recorded but who not receive credit union facilities.

C. Threats to Internal Validity

Threats to internal validity consist of factors which might lead the evaluator to reach incorrect conclusions concerning the cause of the observed outputs of a project or program. While there are many such threats that may arise in each individual project, the most commonly found have been labelled with the following relatively general labels:

(1) History. This term applies to any event occurring during the project which affects the outcome of the project but which is not part of the planned intervention sequence. Any such occurrence is obviously a threat to validity in that the evaluator or manager could not replicate the event elsewhere nor accurately measure its effects apart from those of the project inputs. In Tropicana credit union scenario, the occurrence of drought, flood, or a period of ideal weather during the particular time of the project which results in abnormal crop production would be a threat to validity in that it would prevent an accurate assessment of the credit programs. "History" is perhaps the most obvious and important problem facing the evaluator, expecially since technical assistance projects are usually established without control groups. That is, they are usually established without provisions to monitor selected target groups which receive either no inputs from the project or only select types of project inputs. Drastic events such as revolutions, war, or earthquakes, and even less prominent changes such as the election of a new local government, will have some effect on the administration of a project but will be very difficult to evaluate.

(2) Maturation. Most often occuring in situations utilizing human subjects, this threat involves the mistaking of effects due to the natural processes of development such as aging or maturing, for effects caused by

the treatment administered in the program. In educational projects, for example, increases in the reading ability of students may be due to their "growing up" rather than to the introduction of television into the rural classroom.

(3) Testing. This is a possible threat whenever the target population is interviewed, measured, or tested before and after the introduction of the treatment in the project. It refers to the possibility that the respondents, after taking the first test, will react to the second with the prior experience in mind. For example, the Tropicana farmers, having been asked once about their productivity, may artificially inflate their later responses to the same questions, thinking that they "should" have improved. Such over-reporting would mask the effects of the rural credit program in the region.

(4) Instrumentation. This problem occurs when the method of measuring behaviors or outputs changes during the course of a project. Because new scales and new criteria, or simply more accurate accounting, are introduced one can not compare data collected on an earlier date with later information gathered using the new procedures. For example, if farm income figures were previously calculated on the basis of family farm operations; and the project manager collects income figures on a per capita basis instead, he will have no common scale on which to make a comparison.

Instrumentation is occasionally referred to as the "new police chief effect." When a new, efficient law and order police regime replaces an inefficient or dishonest one, crime statistics in the period after the takeover may show an increase instead of a decrease. This will probably be

due to new or more thorough reporting procedures rather than to increased law breaking activities. Foreign assistance evaluators might also fall victim to this problem. One suspects that often when a health-related or production-related project is implemented, conditions may appear to decline rather than improve due to improved data gathering facilities which are instituted as part of the project itself.

(5) Selection. Better termed as a threat of "self-selection," this pertains to the problems arising when the participants in a program are taken to be representative of the entire population when they are really only representative of a selected minority within that group. If the rural credit program in Tropicana were to extend credit only to those farmers who heard about it on their own, a selection effect would probably be in operation; i.e., the selection of only relatively well-informed and progressive farmers. This would effectively prevent an evaluator from asserting that the results of the project were characteristic of all small farmers within that province.

(6) Mortality. This is in some respects the obverse of the selection problem, because it occurs when different kinds of persons tend to drop out of the treatment groups during the course of a project. It results in treatment populations which are unrepresentative; thus, the real problem facing the evaluator becomes that of discovering why different groups are more receptive than others. One type of foreign assistance project sensitive to this problem is the family planning project, where often the whole point of the project is to discern why certain religious, ethnic, or age groups are more willing to use family planning techniques.

(7) Statistical Regression. This threat to validity is one which has plagued the evaluation of numerous domestic social action programs. Since it is a potential hazard in certain types of foreign assistance projects as well, it is worth taking some time to explain.

Individuals are often selected for participation in educational assistance programs on the basis of extreme scores they make on a pre-test; e.g., either the very best or very poorest students might be chosen for special education programs. Upon examination at a later date it may be found that these individuals' scores are less extreme--the best have become less distinguished and the poorest have apparently improved. These effects, however, are probably the result of "statistical regression," that is, the reversion towards the mean on the second testing of the scores of individuals who were initially extreme. As Carol Weiss suggests, one way to explain this phenomenon is to think of it in terms of measurement error.

> All measures contain some component of "error"....On any one testing, some individuals will score artificially high and others articicially low; on a second testing, their scores are likely to be closer to the mean. If participants and "controls" are chosen on the basis of their extreme scores, they are likely to regress toward the mean, with or without the program. At a second testing, what look like effects of the program may be artifacts of statistical regression.[6]

The evaluator, knowing in advance how individuals have been selected for the program, must proceed with great caution in his judgments. Otherwise, helpful programs such as Headstart might be judged harmful if regression effects are not discounted.[7]

There is no reason to expect that foreign assistance programs might not be influenced also by this effect, although perhaps in quite subtle ways. Choosing a particular region or province for a project either because it has suffered extreme poverty in the last few years or because it has recently become a showcase of development can be seen as analogous choosing students on the basis of extreme scores. Most likely, things will become a bit better in the poorest area and perhaps slip a bit in the best area, i.e. a regression effect toward the mean. In neither of these circumstances will one be able to conduct a fair evaluation of the effects of foreign assistance technologies unless statistical regression is adequately discounted.

(8) Diffusion or imitation of the treatment. This threat to internal validity concerns situations in which the project administration wishes to draw inferences about some larger population based on the results observed in the project. Either a control group is formally built into the project or the project population is assumed to adequately represent the entire population of the region or country. The comparisons may be confounded, however, by diffusion effects which occur when individuals in the program pass on the treatment to the control group. For example, in a foreign assistance project designed to assess the impact of high yield grain varieties, if the farmers in the treatment group distribute their grains to those around them, it will be difficult to draw accurate comparisons between the high yield farmers and the others. This comment does not imply that diffusion is always undesirable; in many projects, such as birth-control dissemination projects, diffusion of information is important. However, in these instances the diffusion process itself is being studied, with the control populations presumably out of range.

A number of other threats to internal validity have been identified in the evaluation literature. These are usually variants of the selection and history threats discussed above, or they are conditions apt to occur only in specialized environments.[8]

D. Threats to External Validity

"External validity refers to the degree of generalizability from one's study to some larger hypothetical population of interest."[9] Rarely is a program or project ocnducted as a one-shot effort without regard for future projects of a similar nature in other areas or further work within the same region or population. Often, in fact, an initial project will be labelled a pilot, prototype, or model program to indicate that it is a trial run for future activities.[10] The degree of external validity, or in other words the representativeness of such a project, will determine the extent to which the decision-maker can make critical inferences about possible applications of the initial findings. There are a variety of threats to external validity; they are usually described as the interactions of "something" with the project activity, that is, to specific conditions within a particular situation which react with the intervention (i.e. the inputs) to create a unique environment.

(1) Interaction of setting and intervention. This problem concerns the extent to which one can generalize beyond the original project conditions to other situations and still expect the observed "cause and effect" relationships to hold. Certainly this threat to validity is crucial to the evaluation of foreign assistance projects. In the Tropicana credit union scenario, for example, the question would be whether or not similar projects would prove as successful if established in other provinces or countries.

The extent of external validity would depend on the degree to which the evaluator felt that conditions had been so unique in the initially-chosen province that they would not be adequately duplicated elsewhere.

(2) Interaction of history and intervention. This concerns the extent to which one feels safe in generalizing about time periods in the past or future on the basis of the results exhibited in a particular program or project. A revolutionary change of national government, or the takeover by a new local village leader, might significantly improve the short-term morale and production of small farmers, but it would make the Tropicana project a very atypical one from which to have to generalize. Thus, history or special events affect evaluation in two different ways: internally in that they might destroy the validity of any cause-effect relationship within the project, and externally in that they might make it impossible to generalize about other projects.

(3) Interaction of selection and intervention. This third matter involves the degree to which the group of individuals involved in a program or project actually are representative of the remainder of the remaining population in that region or nation. If, for example, self-selection took place such that only young, literate farmers participated in the Tropicana credit unions, then the evaluator could not safely draw inferences about the effectiveness of his project in generally raising small farm income within the selected province, or about the effectiveness of similar projects in other provinces or nations. Mortality effects, i.e. people dropping out of programs, could cause a similar threat to external validity.

(4) Interaction of project inputs. This threat occurs when, unknown to the investigator, an individual receives multiple treatments. In a family planning study, for example, a woman who is supposed to be in a

group receiving written information only, might be mistakenly visited by a counselor as well. If this double-treatment occurs frequently, there is no clear basis on which to correctly evaluate the effects of the separate procedures.

(5) Interaction of testing and intervention. This refers to the possibility that the monitoring or testing conditions in a project might be so unusual as to trigger the observed effect, instead of the program treatment. For example, production may increase in the Tropicana province largely because the farmers were investigated so intensively in the beginning that they feld they were under surveillance and consequently took extraordinary care to see that production increased. A similar project tried in another province without extensive monitoring of the participants would not produce comparable impressive results.

E. Threats to Construct Validity

Construct validity concerns the conceptual soundness of the assumed causal relationships operating within the program or project. Threats to construct validity arise when the planning and designing phases of the project are based on assumptions of input-output relationships which do not exist or when the project is implemented in such a way so as to distort the relationships which are to be examined. Four specific threats to construct validity appear relevant to our concerns.

(1) Improper operationalization of the theoretical constructs. The avoidance of this problem is of critical interest to the project designer. Having decided upon the purpose and goals of the planned activity, he must then devise methods of implementing and ways of operationalizing the

concepts involved. If there is no logical connection between the theoretical constructs and the measures employed, construct validity is threatened. In the credit union example, it was assumed that the extension of credit would (a) aid the small farmers in the region, (b) be reflected in increased agricultural efficiency, which (c) could be measured by the amount of grain sold from this region through a central marketing agency. The validity of this project would correspondingly be threatened if the farmers tended to use their credit for consumer purchases, or if they disposed of the grain by means other than the marketing board.

(2) Generalization as to effects over time. This problem refers to the difficulty of accurately specifying in advance the temporal nature of cause-effect relationships in the real world. Inputs or treatments may have several different short and long term effects. The investigator may expect noticeable results within a short period of time, when actually there will be a five year time lag involved; or he may expect any observed effect to be a lasting one, when in reality it will taper off drastically in a short time. Worst of all, he may discover that a program with desireable short term effects has undesireable long term consequences. In situations involving innovative programs, there is really no way of insuring against these possibilities. The evaluator must therefore be alert for unintended program effects and should try to observe the target population for an extended period of time after the initial effects have worn off.

(3) Reactive arrangements. This threat might also be called the Hawthorne effect, a term which has been adopted from a classic industrial productivity study in which a group of workers were monitored to detect the

effect of changed lighting on their output. To their surprise, the investigators found that productivity went up despite the level of lighting; this effect was explained as a reaction by the workers not to the lighting but to the amount of attention they had been receiving.[11] A "reactive arrangement," then, refers to a situation where the treated population is changing its behavior because of the arrangements of the study rather than because of the delivered input. Certainly, in the credit union example farmer's production may have increased merely because of the attention they were receiving from foreign or governmental officials.

(4) Interaction of participants and inputs. This problem might be considered as a facet of the reactive arrangements threat, but it is interesting enough to deserve separate mention. Occasionally the respondents in a program will learn enough about the setup of the study to alter their behavior, not necessarily according to the treatment but instead according to the time period of the program or the variety of assistance which is available. Cook and Campbell cite the example of participants in the New Jersey Income Maintenance Experiment who having learned that they were guaranteed a fixed minimum income for three years, conditioned their behavior in light of this three-year expectancy.[12] In the hypothetical credit union project, farmers might adjust their credit application activities and their loan repayment activities according to the period of time they felt foreign or government officials would remain on the scene.[13]

F. Summary

This long list of threats to validity may be somewhat discouraging. Certainly anyone who has been involved in any project or program evaluations will recognize that one or more of these pitfalls existed in each and every

instance. The perfect project and the perfect evaluation are essentially figments of the administrator's imagination. Nature simply does not cooperate to the extent that all threats to validity may be dismissed. However, the purpose of the list was not to serve as discouragement to the evaluator but rather as encouragement and challenge. Armed with knowledge as to the possible problems which may arise, the decision maker may take steps to offset as many threats as possible by planning and implementing a project design and evaluation which avoids or circumvents potential pitfalls.

The next section of this study, therefore, is devoted to describing a series of designs and exposing their individual strengths and weaknesses. First to be discussed is the so-called experimental design which succeeds in avoiding many of the threats to validity. However, this type of design sees limited application in social action programs, especially in foreign assistance programs. We turn next to an extended treatment of quasi-experimental designs. These are more adaptible and thus are commonly employed in evaluation studies. Each type of quasi-experimental design has separate advantages and disadvantages, i.e. it is susceptible to some threats to validity but immune to others. Our discussion will therefore focus on exposing those characteristics of each design which make more or less suitable in the context of foreign assistance project evaluation.

The admission that there exists no perfect design, however, leads to another question, namely which of the threats to validity are of most importance? Since compromise is necessary where should it come--at the expense of internal validity, external validity etc.? Should the investigator be more worried about the generalizability of the findings (external validity) or the exactness of the input-output or cause and effect relationship

exhibited in his findings (internal validity)? The answer, of course, is that it depends. It depends on the type and purpose of activity is being evaluated. One single project cannot answer all of the administrator's questions or avoid all of the threats to validity.

Evaluators generally argue that despite the level and purpose of an investigation internal validity should be of primary concern, with trade-offs occuring between external and construct validity and finally statistical validity. Their reasoning is that internal validity determines whether or not there is valid causal relationship demonstrated in the program being studied. If there is not, or if threats to internal validity are severe enough to cause serious doubt, then there would be no point in continuing further with the venture. After internal validity has been established then other threats to validity become important. Two situations especially appear typical of those encountered by foreign assistance evaluators in AID.

First, there is the type of investigation concerned with whether or not a particular causal relationship holds in certain developing nation environments; for example, whether or not an increase in the supply of capital available to the small farmer will ultimately help alleviate rural poverty. Concern here is with concepts of a very broad nature, that is with construct validity. The priority ordering in studies of this kind will be internal validity, construct validity, statistical conclusion validity, and external validity; because for the moment the investigator is concerned with causal relationships and conceptual matters. AID participates in ventures such as this, however, to a decreasing extent. The agency tends to shy away from "research" studies and administers very few pilot or experimental

projects. Although this perspective may be slightly altered in light of new congressional mandates; AID essentially sees its role in the area of applied action programs.

The priorities for this second kind of investigation are understandably quite different. Decision-makers, particularly those who must account for and justify the expenditure of valuable resources, are interested in what works and where it can be used, that is, effectiveness, cost-effectiveness, and impact. This translates into a concern first with internal validity and secondly with external validity, because once having found a "successful" formula, the administrator looks for other places where he can apply the same solution. In our view, this is a characterization which most applies to AID and its interests when evaluating foreign assistance projects. Therefore, in delineating various research designs, we shall attempt to highlight those which maximize aspects of internal and external validity.

NOTES

[1]James Caparaso, "Quasi-Experimental Approaches to Social Science," in his Quasi-Experimental Approaches: Testing Theory and Evaluating Policy, Evanston, Ill.: Northwestern University Press, p. 8.

[2]Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research, Chicago: Rand McNally and Co., 1966 is perhaps the most prominent work. See also Donald T. Campbell, "Reforms as Experiments," The American Psychologist, vol. 24, no. 4, April, 1969, pp. 409-429; and Thomas D. Cook and Donald T. Campbell, "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings," in M.D. Dunnette, ed., Handbook of Industrial and Organizational Research, Chicago: Rand McNally, 1975.

[3]Thomas Cook and Donald Campbell, op.cit., pp. 5-10.

[4]Ibid., p. 20.

[5]See ibid., pp. 14-19.

[6]Carol B. Weiss, Evaluation Research, Englewood Cliffs: Prentice-Hall, Inc., 1972, p. 70.

[7]For a discussion of the Headstart program evaluation refer to: Donald T. Campbell and A. Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth, ed., Compensatory Education: A National Debate, vol. 3 of Disadvantaged Child, New York: Brunner-Mazel, 1970.

[8]Thomas Cook and Donald Campbell, op.cit., pp. 11-12.

[9]Ilene Bernstein, George Bohrnstadt, and Edgar Borgatta, "External Validity and Evaluation Research," Sociological Methods and Research, vol. 4, no. 1, August 1975, p. 102. This article presents a detailed analysis of several of the major threats to external validity.

[10]See Edward A. Suchman, "Action for What? A Critique of Evaluative Research," in Carol B. Weiss, ed., Evaluating Social Action Programs: Readings in Social Action and Evaluation, Boston: Allyn and Bacon, pp. 60-61.

11

[12]Thomas Cook and Donald Campbell, op.cit., p. 28.

[13]Cook and Campbell, ibid., cite several other threats to construct validity as well, pp. 25-29.

Chapter III

Experimental and Quasi-Experimental Designs

A. The Principles of Experimental Design

Concern for sound evaluation in light of the many threats to validity has led experts to advocate the experimental model as the soundest type of research design. This is because of the rigorous and controlled procedures employed by the investigator, often within the confines of a laboratory. To the extent, however, that we will deal with experimental designs, we are concerned with non-laboratory situations. These are sometimes called field experiments. As Cook and Campbell comment, "a 'field' constitutes any setting which respondents do not perceive to have been set up for the primary purpose of conducting research."[1] That is, a field experiment is a program involving a natural setting.

These two authors go on to define an experiment as

> any experimenter-controlled or naturally occurring event (a treatment) which intervenes in the lives of respondents so that its consequences can be empirically assessed and a cause-effect relationship collaroborated.[2]

For our purposes, it will be useful to cite and analyze the somewhat more detailed description of an experiment given by Reichen and Boruch in their volume on social experimentation.

> By experiment is meant that one or more treatments (programs) are administered to some set of persons (or other units) drawn at random from a specified population; and that observations (measurements) are made to learn how (or how much) some relevant aspect of their behavior following treatment differs from like behavior on the part of an untreated or control group also drawn at random from the same population.[3]

Their definition serves to highlight seven basic aspects which are involved in the ideal experimental design. Of central importance are (1) dependent variables, the conditions or factors under study which are to

be altered in some manner; (2) independent variables, the conditions which are thought to be related to, or "cause" changes in the dependent variables; (3) a target population, a specified population of units in which the relationships between the independent and dependent variables is to be studied (Note that the units need not be humans, but may be fields, villages, provinces, etc.); and (4) confounding variables, factor which may obscure, alter, or interfere with the presumed cause-effect relationship in the study. Confounding variables may be seen as causing the threats to validity discussed earlier. Experimental procedures then involve (5) manipulation of the independent variables by the experimenter in order to cause some change in the dependent variables; (6) measurement, before and after the intervention, so that the extent of the effects may be accurately assessed; and (7) control of the confounding variables so that spurious relationships are avoided. Control, which is crucial to the ideal experiment, is implemented in two ways: (a) by the division of the subjects into a treatment group and a control group or groups which do not experience the effects of the manipulation but which are still observed and measured along with the treatment group; and (b) by a procedure of random selection and assignment of individual units into these groups. Further discussion about the rationale for these control procedures may be helpful.

As was pointed out in the previous chapter, the establishment that an intended causal relationship existed within a project hinges upon the investigator's ability to show that something happened to the units receiving the inputs which did not happen to all other similar units. In the Tropicana example, the investigator has to show that any productivity increases for farmers involved in the credit union project were not typical of all the other

farmers in the province or nation who were not involved in the credit unions. Demonstrations of this nature are facilitated by setting up control groups, i.e. groups of units, as similar as possible to the treatment group, which do not receive the same inputs as the treatment group but which are observed and measured in the same fashion as the treatment group. It should be emphasized that being in a control group does not necessarily mean that these units receive a "lack of treatment," that is, no inputs at all. A control group may receive a different variety or different degree of input and still serve as a comparison to the full treatment group. Often, for example, family planning projects will be established in which the target population is divided into several groups. Each group will then be given a different type of birth control device and/or a different regimen of counselling and visitation. Evaluation, therefore, involves comparison among the results of these different groups to determine which program of inputs had the most impact or was most cost-effective. In many writings, particularly those dealing with so-called quasi-experimental designs, the term comparison group(s) is used instead of control group(s). This indicates to the reader that while a group or groups have been separated from the treatment group for purposes of comparison, these group(s) were not selected or assigned on a random basis. Usually what has happened is that the project designer has tried to "match" certain characteristics of the groups and has assigned units to groups on these grounds.

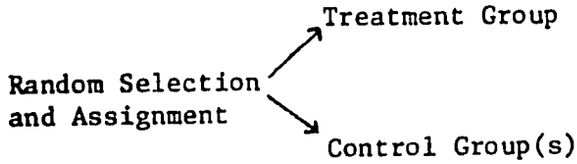
The key to establishing control or comparison groups is that they be chosen to be as similar as possible to the treatment group. Ideally, individuals in different groups will be distinguishable only on the basis of the varied inputs which they received in the project. Selection of a target population as a whole and division of this population into groups is thus of

prime importance. The most efficient method for performing these tasks is through randomization, i.e. random selection of individuals for involvement in the project or program and random assignment of individuals into the control groups. Any large population of individuals may be distinguished on numerous salient characteristics, e.g. weight, height, size, age, color, income, production, etc. Random selection of a sample from among such a population is the only sure method of obtaining a group of individuals which accurately reflect all of the features of their parent population. Random assignment further assures that individuals within control groups will be similarly alike.

Evaluation under ideal experimental conditions is relatively straightforward. The basic notions are illustrated in Figure III.[4]

Figure III

Basic Features of Experimental Design



	Measurement	
	Before Intervention	After Intervention
Treatment Group	a	b
Control Group(s)	c	d

Indicators of the dependent variable are measured in the control and treatment groups before and after administration of the inputs (indicated as a, b, c and d in the figure). Comparisons are then made (1) within groups, (a vs. b, c vs. d), to determine if any change took place, and (2) between groups,

(a-b vs. c-d), to determine whether or not significant change might be attributed to the treatment. If the difference between the treatment group indicators is greater than the differences between the control group indicators, a true effect has been established. The significance of this effect can be further analyzed using statistical procedures. Randomization has assured that the two comparisons are valid by providing groups which are representative of the population being studied (i.e. through random selection) and by further providing groups which are comparable in all characteristics except the treatment (i.e. through random assignment.)

The hypothetical Tropicana credit union project could have been designed to follow an experimental format as follows: First, all rural villages within the province would be enumerated and a portion chosen randomly for inclusion in the project. These villages then would be randomly assigned to treatment groups and to control group(s). Measurements would be taken on indicators of productivity and income of the small farmers in each of these villages. The credit unions would then be established and a series of measurements over time would be taken on the treatment group and control group farmers. Evaluation and assessment would then proceed with this data. It should be noted that random selection in this example did not involve individual farmers. It would be impractical to attempt to establish credit unions in villages and then to allow only particular farmers in those villages to be members. Assuming that the characteristics of the rural villages were alike, adequate comparisons could be made when the village was utilized as the selection unit.

Experiments administered according to the procedures outlined above would constitute very "sound" designs for evaluation--sound in the sense of circumventing threats to validity. In fact, the experimental design effectively

thwarts all internal threats to validity. External validity is assured to the extent that the treatment and control groups were chosen from a relevant population. (Construct validity remains a matter of conceptual design.) However, despite these strengths, the use of experimental designs in social action programs, domestic and foreign, is much more widely advocated than it is practised. There are various reasons given for this, most centering upon the issues of randomization and control groups, the cost of establishing such procedures, and the problem of generalizing from "experimental" designs.

Thus in many instances, the decision maker does not employ randomization in selecting and assigning units to groups within projects or programs. Usually he will argue that random selection was too expensive; since to be effective, careful, advance enumeration and monitoring of the population is required and a great deal of work is involved in drawing a good sample. Or, alternately, randomization may be opposed for political or normative reasons, e.g. people should be selected only on the basis of need. The idea of control groups similarly, though mistakenly as we have pointed out, suggests to some decision-makers the denial of services and thus these too are often ommitted in project and program design. The amount of supervision, data collection, and analysis required in an experimental design is regarded as too expensive, too time-consuming, and too sophisticated to be worth the effort. Many of these arguments, however, are based on quite exaggerated claims and/or misconceptions as to what experimental design really involves.

Others, however, argue on methodological grounds that experimental designs are not totally appropriate for evaluation purposes. They feel that the experimental model is suited for a single-treatment, one-shot, effects-only

analysis. The logic and practicability of on-going evaluation of large-scale social programs supposedly precludes implementation of such controls and procedures.[5] Furthermore, external validity is likely to be threatened by the isolation, attention, and recurrent measurement procedures entailed in an experimental design.

These objections are not universally accepted. Proponents of the experimental method argue that if enough care were shown in design and selection, that if budgets were realistically appropriated, and that if people really understood the meaning and necessity of randomization procedures, then a great deal more truly experimental work would be done. Reichen and Boruch in their Social Experimentation volume and in their 1974 report to AID recommend using experimental methodologies in the design and evaluation of foreign assistance projects.[6] These authors make a strong case for the appropriateness of randomization in situations: (a) where there is not an adequate supply of the "treatment" to satisfy the demands of the entire population, (b) where a program is being tried on a trial or innovative basis, or (c) when it is important that program effects be carefully monitored and evaluated.

B. Quasi-Experimental Design: A Definition

Regardless of the reasons, be they practical, financial, or logical, it is clear that experimental designs are not widely utilized in evaluating social action programs.[7] This raises the question of alternatives. The fact that all aspects of an ideal model cannot be implemented does not mean that the feasible concepts should not be applied. In the fields of research design and evaluation, there has developed a series of alternative techniques and procedures to ensure systematic investigation and evaluation under less

than optimal conditions. These have been labelled quasi-experimental methods. While initially referring to a very broad range of topics, the term now usually refers to a school of thinking about research design and evaluation due to the efforts of Donald Campbell, whose recognition of the problems of evaluation in social science led him to formulate and systematize a series of methods which could be utilized in non-experimental situations.[8] The remaining portion of this section will be devoted to an explanation of a variety of such different designs for evaluation purposes, together with a look at the strengths and weaknesses of each.

First of all, however, a few words of clarification about the phrase "quasi-experimental" are in order. Presumably this term could be used loosely to describe any design or evaluation which is "non-experimental" in nature, i.e. lacking one or more of the seven aspects of a true field experiment. This, however, is a bit too broad a notion to cover what is usually referred to as a quasi-experimental design. Instead, the distinguishing feature of this type of design is the absence of randomization procedures in the selection of units and assignment to control and treatment groups. In many other regards the quasi-experimental situation usually approaches the experimental. As Reicken and Boruch state:

> Quasi-experiments correspond to certain natural social settings in which the experimenter can approximate experimental procedures for collecting data even though he lacks full control over the delivery of the treatment.[9]

Quasi-experimental designs, therefore, still use the concepts of independent and dependent variables, the measurement of change, and the manipulation or administration of treatment to a target population. They fundamentally differ from true experiments in the control over extraneous factors. In all

the quasi-experimental designs considered below randomization is absent, and in some of the designs there are no comparison groups. Obviously, threats to validity are of critical concern to the evaluator in these circumstances. Some quasi-experimental designs prove to be much stronger than others in this regard, as will be seen below.

C. A Selection of Quasi-Experimental Designs

Six different types of designs utilized in social action programs and evaluations will be presented. They are arranged in order of increasing complexity and sophistication, and to a large extent in order of increasing soundness, in that threats to validity are avoided. The first three designs are "weak"--most evaluators would not grace them with the label quasi-experimental. However, these designs are frequently used, so their strengths and weaknesses should be illustrated. Furthermore, these initial formulations provide a good introduction to the more legitimate quasi-experimental designs. As will be seen, in some cases often a relatively minor modification of a weaker design can transform it into a much sounder design.

Each quasi-experimental design will be illustrated by a schematic diagram. The following terms are used in conjunction with these diagrams:

Pre-test: the measuring or testing of the populations involved prior to administration of the project or program inputs. Not all the designs have pre-tests.

Post-test: the measuring or testing of the populations involved after the administration of the project or program inputs. All the designs have post-tests.

Single-group: a design which does not involve use of a comparison group. The inputs or treatment is administered to all the individuals in the program or project population.

Comparison Group: a subset of the population that is observed and measured but which does not receive the same schedule of inputs as the treatment group. Comparison groups serve the same function in quasi-experimental designs as control groups do in experimental designs. The difference is that random selection and assignment are used to set up experimental control groups. Quasi-experimental comparison groups are established by nonrandom assignment by the investigator, who attempts to match the characteristics of the groups on selected salient criteria, e.g. age, wealth, race, religion. Since randomization is the only sure method of ensuring the equivalence of groups, the phrase non-equivalent control groups is occasionally used in the quasi-experimental literature to refer to what we are calling comparison groups.

Time series: a series of observations made of the population, usually taken at regular intervals, before, during, and after the administration of the project or program.

The following symbols are utilized in the diagrams: 0 represents an observation or measurement, either a pre- or post-test; X represents the administration of the inputs of the program. Observations and treatments given to a particular group are recorded on one line, with an implied temporal ordering from left to right. 0's recorded directly above or below each other indicate simultaneous measurement of the groups involved. Figure IVa, therefore illustrates a design with a treatment group (top line) and a comparison group (bottom line) with observations (the 0's) taken of each group before and after the administration of the program (X).

Figure IV
Two Examples of Program Design

0	X	0
0		0

IVa

0	0	X	0	0

IVb

On the other hand, Figure IVb illustrates a single group design in which time-series measurements have been taken, i.e. two pre-test observations and two post-test observations.

A word is also in order concerning our usage of the terms pre-test, post-test, and treatment. These are terms most commonly used by laboratory researchers and clinicians. Some readers may therefore find their association with quasi-experimental designs, particularly in the context of foreign assistance project evaluation, a bit out of place. Our usage is essentially one of convenience: it is rather awkward to substitute "pre-project observations" for pre-tests, or "post-program measurements" for post-tests, or "administration of project inputs" for treatment. Thus we will continue using the shorter phrases with the intention that they are to be applied to their counter-parts in a quasi-experimental environment.

(1) <u>The Single-Group, Post-test Only Design</u>. This design has also been called the <u>one-shot case study</u> for reasons which are apparent in Figure V.

X	0

Figure V
The Single-Group Post-test Only Design

In many ways it is charitable to label this a design at all. It essentially consists of a single observation of the treatment group sometime after the program has been administered. Since there is no pre-treatment observation or any comparison group(s), comparisons within groups or between groups are not possible. Any evaluation done in this type of situation will thus be impressionistic and narrative in nature. None of the threats to validity, internal, external, conclusion, or statistical, can be effectively eliminated in this design. There is no conclusive way to demonstrate any relationship, to assess its direction, or to determine its significance.

This is the weakest form of evaluation, and hopefully one which is seldom utilized by administrators of social action programs. Certain circumstances may dictate its usage, however, such as (a) the absence of any funds or personnel for systematic evaluation, or (b) the inability of the administrator to control the treatment or effect he wishes to observe. This latter situation could occur, for example, if one had to recount the effects of flood relief or disaster assistance. As it would have been impossible to predict the occurrence of such events, "pre-tests" to establish the state of prior conditions do not exist. Evaluation, therefore, is confined to a reporting of the delivery of the inputs, e.g. the number of vaccinations given, the amount of food supplies received etc.

Suppose, however, that a decision-maker was called upon to evaluate the impact of the introduction of high-yield varieties of rice in Pakistan in the 1967/68 crop year.* If the only information available to him was the fact that rice yields in that single year were 16 quintals per hectare, there

---

*This figure and all subsequent figures utilized in conjunction with the high-yield varietal example are taken from a recent USDA-AID publication entitled Measuring the Green Revolution: The Impact of Research on Wheat and Rice Production, authored by Dana G. Dalrymple, Economic Research Service, USDA Foreign Agricultural Economic Report No. 106, June 1975.

is really nothing that he could say on the matter. Without base line data on yields prior to the program there is no basis for comparison. Although this example is somewhat simplified, it serves to illustrate the overwhelming drawback of the post-test only design. Meaningful evaluation is impossible. As we progress to other designs we shall see how a much sounder evaluation may be achieved through the collection of additional data and the use of comparison groups.

(2) The Single Group Pre- and Post-test Design. As Figure VI indicates this format involves the addition of pre-treatment observation to the previous design. The individuals in the program are observed before and after the delivery of the inputs, allowing for within group comparisons to determine

0	X	0

Figure VI
The Single Group Pre- and Post-test Design

the extent and direction of any change. This design is obviously a significant improvement over the post-test only design; however, it does not allow for between-group comparisons and becomes susceptible to a number of threats to validity, particularly internal validity.

The Tropicana credit union example, can be used to illustrate the problems which might arise with internal validity. Let us presume that small farmer production increased from 15 to 25 bushels per acre, as evidenced in observations made before and sometime after the establishment of the credit union program. Although crop yields increased over 60 percent,

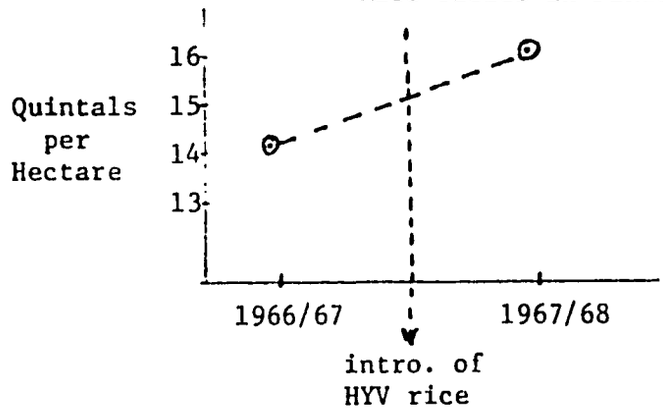
this may have been because of ideal growing conditions which in fact resulted in even larger increases in other provinces in Tropicana--an example of the threat of history. Selection or mortality may have been a problem in that only the larger, more efficient farmers, instead of the smaller, poorer ones, utilized the credit facilities. Testing or instrumentation might have jeopardized the results because new and thorough measurement procedures may have been introduced in conjunction with the program. The use of a comparison group, i.e. the farmers in villages which did receive credit unions, could have prevented most of these threats to internal validity.

External validity also remains a concern in this design. Because the program is administered only to a single unit where the individuals are not assigned randomly, the generalizability of the results to other units or populations is in question. For example, if the particular province in Tropicana were chosen as a showcase province in which the rural population sympathized with the central government and foreign administrators, inferences about similar successes in other regions of the country may be unfounded.

Consider also the example of evaluating high-yield varietals in Pakistan. A pre-test, post-test design would look like this:

Figure VII

Rice Yields in Pakistan



The evaluator is now in a position to make a statement as to the change in production from the year before to the first year after the introduction of the new varieties. That is, he can show that yields increased about 10% from 14.5 to 16.0 quintals per hectare. Various confounding factors, however, must be eliminated before he can confidently attribute this increase to the new varietals. For example, perhaps the increase was due to good weather, perhaps 1966/67 had been an exceptionally bad year, or perhaps agricultural census procedures were substantially overhauled to account for previously neglected production areas. Many of these threats to validity can be overcome with the adoption of more stringent design procedures.

The single group pre- and post-test design is frequently used to evaluate social action programs. It endangers, however, internal and external validity because of the absence of a control group and the non-random nature of the selection. The usually short interval of time between pre- and post-test measurement also means that any observed change may be an artifact of seasonal or short-range changes in the environment rather than an indication of a long-term pattern or effect. All of these faults may be remedied to a large extent by more elaborate research designs. However, in instances where they are not feasible, evaluators can utilize this design and draw useful conclusions from it provided that they are aware of its drawbacks.

(3) The Post-test Only Design with Comparison Group. This design is also called the static group comparison design. It includes a non-equivalent control group but does not involve a pre-test. As Figure VIII indicates, observations are made on both the treatment and comparison groups, but only after the administration of the program. We shall assume for the moment

Figure VIII

Post-test Only Design with Comparison Group

X	0
	0

that it is not possible to obtain pre-program information through archives or census data. This design is therefore particularly susceptible to two threats to validity: selection and treatment. The first concerns whether or not the comparison group is equivalent to the treatment group. Without any prior information about the comparison and treatment units, there is no obvious basis on which to make such an assumption.

If these circumstances applied to the hypothetical Tropicana case, for example, the investigator would have agricultural yield figures for the treatment province and one other province for a crop year following the establishment of the rural credit union program. He would then be forced to try and draw comparisons without knowing if one or both of the provinces were perhaps the most fertile in the country, or the most barren. He cannot assume, therefore, that some sort of selection bias is in operation, or that the observed treatment effect is not an artifact.

Even if the investigator were reasonably confident about the equivalency of the two groups in the design, there would be no way of ensuring that any observed difference is a program effect and not a "treatment" or "testing" effect. The Tropicana farmers could have responded solely to the conditions or measurements associated with the program, or to the attention they were receiving, instead of to the increased availability of credit. Other threats

to validity might arise because this design is dependent upon observations taken at a single point in time. There is no way to eliminate problems of history, for example, or to discount maturation, or to account for long or short-term trends. All in all, if only post-test information is available, this design does not allow the evaluator to draw any conclusions about the success or the causal relationships in a program.

Unfortunately, AID administrators and field officers face this problem on many projects. Adequate data for developing countries, particularly at disaggregated levels, does not always exist. Often part of the project itself is the establishment of data-gathering and recording facilities. The location of a technical assistance project may be arbitrarily chosen for special reasons, such as political considerations or natural disasters; even though the pre-test information is not available the program must be implemented by a certain date. A comparison group may be also arbitrarily chosen as the "showcase" village or the closest village. Under such circumstances, the administrator and evaluator have little freedom of action. The best they can do is use informal evidence, archives, or post-test observations to establish (a) the state of conditions in the treatment group prior to the program, and (b) the extent of the equivalence between the control and treatment groups. The degree of soundness of any evaluation made in a project using this design depends on the extent to which the evaluator can draw informal comparisons within and between the groups.

If some additional information is available, however, either about the population as a whole or about cohorts of the individuals involved in the program, there are versions of the post-test only design which are somewhat stronger and which could be adapted for use in AID project and program evaluations.

(a) A Multiple Comparison Group Post-test Only Design. This special design requires that the investigator obtain data for the entire population, treated and untreated, on both the chosen post-test criteria and some other quantifiable dimension that is presumably related to the assignment of treatment. By arraying the post-test results against this second dimension, it may be possible to detect discrepancies between the treated and untreated segments of the population that may be attributed to the effects of the program.[10] An actual example of these procedures probably will make things much clearer.

In 1972 Fleiss and Tanur were interested in discovering if the recipients of Medicaid (a U.S. federal program to subsidize medical care to the poor) were availing themselves of their opportunities through the program to obtain medical assistance.[11] Eligibility to this program was restricted to individuals making less than $3000 per year. Since Fleiss and Tanur were limited to post-test only information, they could not compare the average number of visits to a physician made by people before and then after they had entered the program. So instead, the evaluators plotted the number of physician visits, (measured after the program was implemented), against annual income figures for the entire population, both recipients and non-recipients. The results are indicated in Figure IX. It is apparent

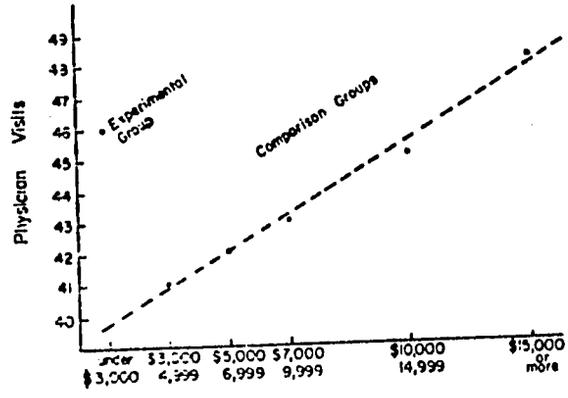


Figure IX

An Example of the Post-test Only Multiple Comparison Group Design (taken from Reichen and Boruch, Social Experimentation, p. 115)

that the number of visits to a physician was directly and positively related to individual income, i.e. the more income, the more visits. The one notable discrepancy was the income group of less than $3000; the Medicaid recipients visited doctors much more frequently than persons in the immediately higher income groups. If it is presumed that use of medical advice is restricted by income, and that Medicaid recipients tended to use medical facilities more often because these restrictions were removed, this diagram and this design indicate in a striking manner the effects of the program.

Applications of this relatively simple but effective evaluation design might be carried out in the foreign assistance field as well. For instance, if the Tropicana credit union loans were available only to those farmers whose holdings were less than some specified average, if previously in this province yields were directly related to the size of the farm, and if post-test data were availabel on production yields for all farmers in the province, then a diagram similar to Figure X might be constructed. Results such as those drawn in the Figure would provide strong evidence of the successful impact of the project.

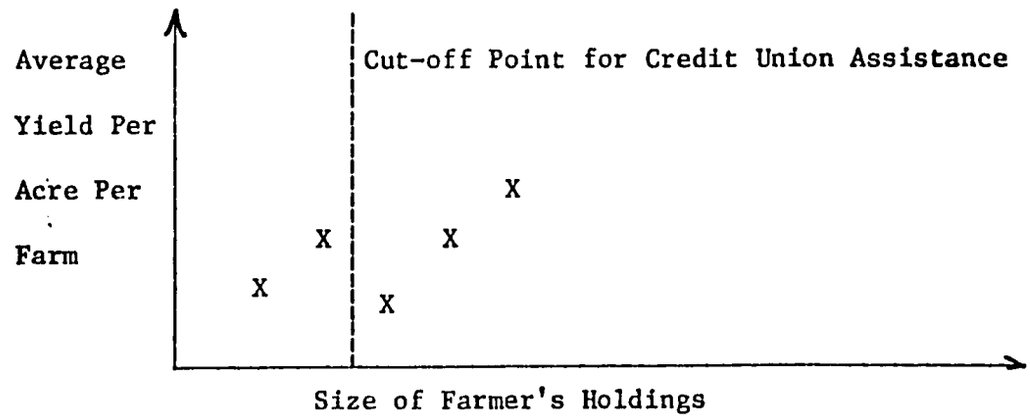


Figure X

Design Example for the Hypothetical Tropicana Credit Union Project

Given the circumstances which confront the technical assistance officer, particularly the paucity of adequate pre-program data, this type of post-test only quasi-experimental design might offer a partial solution to some evaluation problems.

(b) Cohort Group Designs. The format of this special design is unchanged from that indicated in Figure VIII: post-program measurement of both control and comparison groups. It is, however, particularly appropriate for situations involving human participants such as training or educational programs. In these circumstances, a comparison group of "cohorts" may be selected, i.e. a group of persons who have the same backgrounds, ages, incomes, and other characteristics as those in the program, but who have not yet or who will not enter it. Measurements taken on the cohort group are substituted by the evaluator for pre-program measurements of the program group itself. Assessments of the effects of the program are made by comparing the cohort and post-program group indicators. The soundness of this type of evaluation is totally dependent upon the degree of equivalency of the cohort group to the program group. As we emphasized before, random selection and assignment is the only sure way to ensure equivalent groups. Designs involving selected cohort groups, therefore, should not be regarded as substitutes for designs involving randomization. The former should be employed only if circumstances preclude the latter, e.g. the program is already underway when the evaluator is consulted.

Consider, for example, a technical assistance program which has been already set up to teach mathematics to school children through a series of radio broadcasts. Cohort groups of students could be selected and used as comparison groups to determine the efficiency of the new techniques.

Presumably these youngsters would be of the same age, locale, and experience and would not be chosen because of any unusual characteristics such as intelligence. A variety of possible procedures could then be adopted.

(i) When the program is over, the treatment group and a cohort group of students about to enter the program could be tested. The evaluator must be cautious, however, because observed changes could result from maturation of the students while in the program, or testing effects, instead of the result of the broadcasts. This evaluation format is described in the diagram below:

X 0	program group
0	cohort group

Figure XA
Cohort Group Design, Example One

(ii) When the program is over, the treatment group test results could be compared to those of a cohort group whose members had been tested in the past under the old methods of instruction. Here the individuals in the two groups would be of the same age, maturity etc. when tested, but threats to validity would arise from history. This design could be diagramed as follows:[12]

0	cohort group
X 0	program group

Figure XB
Cohort Group Design, Example Two

(iii) When the program is over, the treatment group and a cohort group of students who simultaneously received the old instruction methods could be tested and compared. This would be the strongest of the three designs because the threats of maturation, selection, and history present fewer problems and the evaluator can obtain a good indication of the differential effects of the new and old programs. However, such procedures would involve either segregating one group of students from another, which might be practically impossible due to facilities or diffusion effects, or selecting the comparison group from a different school or locale, in which case the cohort characteristics of the group may be less equivalent.

All of these three cohort designs depend on the evaluator's ability to find a cohort comparison group as similar to the treatment group as possible. Such designs have been utilized in foreign assistance projects, particularly those in the medical and educational areas.[13] While their usage is certainly to be encouraged in circumstances where random assignment and selection are not possible, it must be remembered that only through randomization can truly equivalent groups, and thus the soundest evaluation, be attained.

(4) <u>The Pretest, Posttest Design with Comparison Group</u>. This design is similar to the true experimental design in that it involves pre and post program measurements on both treatment and comparison groups. However, since the selection and assignment of these groups is not done randomly, the "control" group is nonequivalent. Figure XI indicates the format of the two-group pretest, posttest design. This design controls for most threats to

0	X	0
0		0

Figure XI

Pretest, Posttest Design with Comparison Group

internal validity, although problems of selection, maturation, statistical regression, or local history could occur due to unrecognized nonequivalency of the comparison and treatment groups. Selection-maturation effects, for instance, occur when one group of individuals is brighter, more motivated, more mature, or older than the other. Thus, between-group differences may be mistakenly attributed to program effects when they are in reality reflections of some undetected process of growth or change in the respondents.

The use of a comparison group should remove the threat of history. However, because the comparison group in this design is not randomly selected from the same population as the treatment group, it is possible that "local history," special events restricted to a particular region, village, or school, might upset the validity of the evaluation. In the Tropicana credit union example, if one province were used as a comparison while the project was established in another, a change of provincial governments or an incidence of local flooding could make between-group comparisons faulty.

The pretest, posttest type of design is one which is certainly relevant to the foreign assistance project evaluator. Its format allows rather clear-cut comparisons and if comparison groups are chosen with care, the evaluator can be reasonably sure of having avoided most threats to validity. Once again, however, the possibility of statistical regression effects should not be discounted; projects which are established in "showcase" villages or in areas affected by extreme weather conditions, for example, would be highly susceptible to this problem.

While the foreign assistance officer will usually be satisfied if he can obtain good pre- and post-program measurements, this data base does not ensure external validity concerning long term effects and cycles or

differential growth rates. Figure XI, which is constructed to illustrate data which might come from a situation such as the Tropicana project highlights these possible problems. The results attributed to the

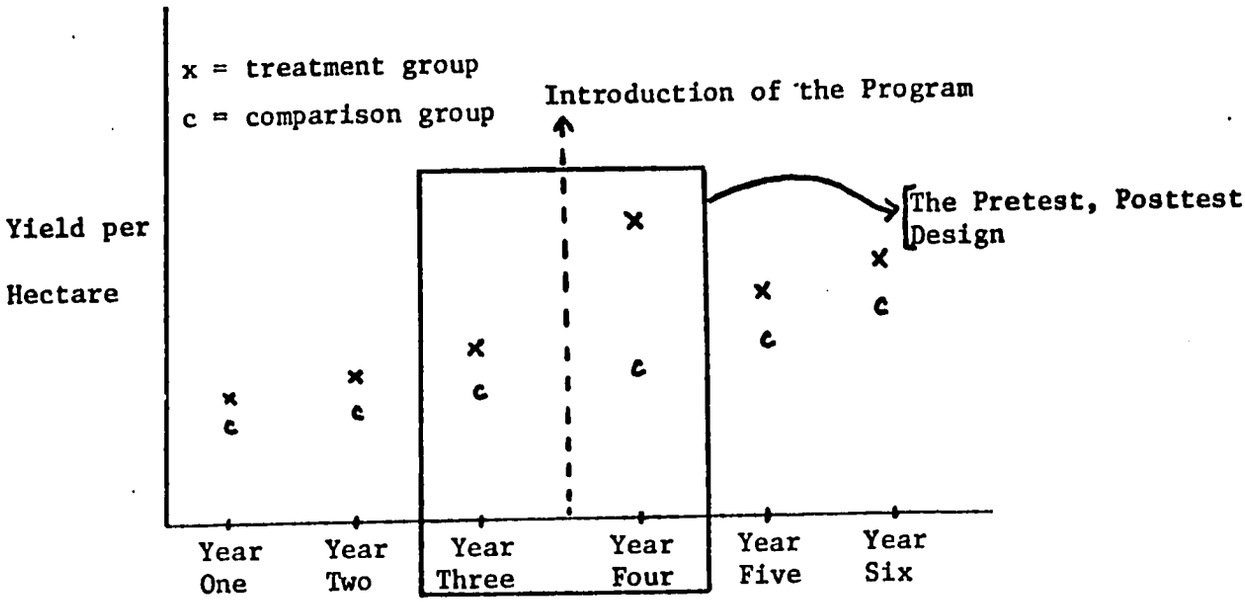


Figure XI
An Example of Possible Problems in the Pretest, Posttest Design

program may actually be only a reflection of immediate short-term fluctuations in yields which happened to coincide with the implementation of a program. As the boxed-in area in the Figure indicates, an evaluation constructed using only those observations could reach conclusions that would be unsupported by subsequent events. Time series designs, the fifth and sixth types of quasi-experimental formats to be discussed in our catalogue, are utilized to overcome such difficulties.

Returning to the pretest, posttest design for a moment, Cook and Campbell indicate a variety of outcome patterns which may occur in this design.[14] These are sketched in Figure XII. The investigator should determine which situation

Figure XII about here

Figure XII

Five Possible Outcomes of the
Pretest, Posttest Design

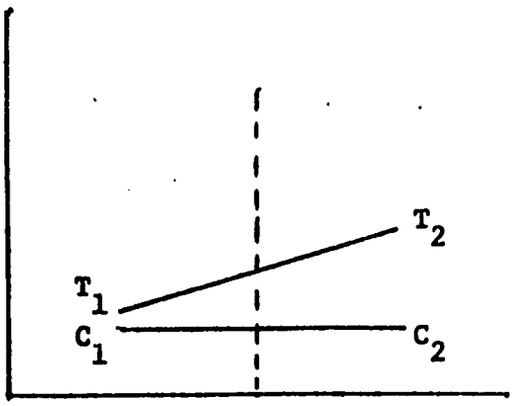


Figure XIIA

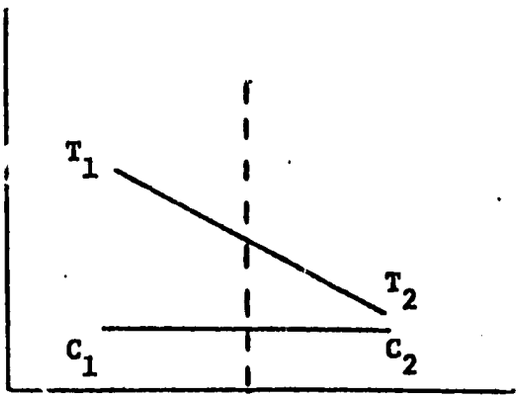


Figure XIIB

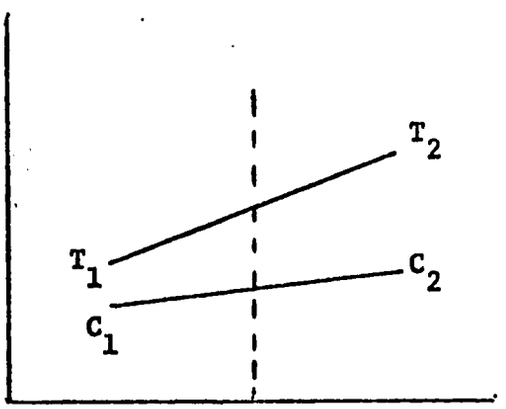


Figure XIIC

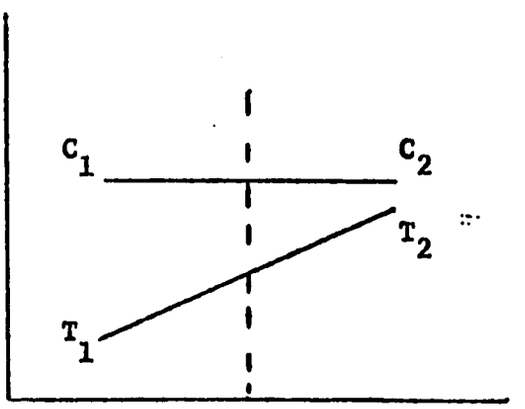


Figure XIID

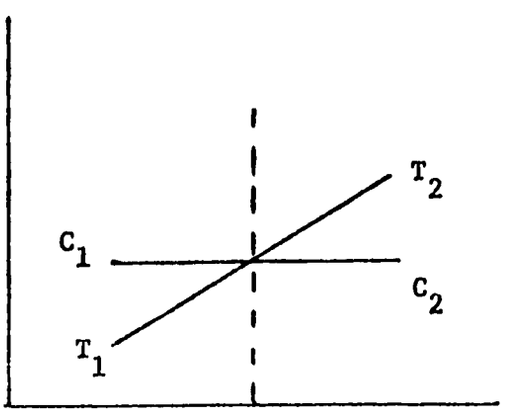


Figure XIIE

T_1 = Treatment Gp., Pretest

T_2 = Treatment Gp., Posttest

C_1 = Comparison Gp., Pretest

C_2 = Comparison Gp., Posttest

is relevant to his concerns prior to conducting the evaluation; then, if the results do not conform to expectations, he can search for extenuating circumstances. In Figure XIIA comparison and treatment groups are similar at the start of the program and it is expected that a noticeable change will take place in the treatment group but not in the comparison group. Figure XIIB, on the other hand, indicates a situation in which the program is expected to alleviate or depress some condition, such as incidence of disease, while the comparison group, having undergone treatment before, should remain unchanged. Here one is looking for the treatment group to become similar to the comparison group. In Figure XIIC differential growth rates for the treatment and comparison groups will make the outcomes difficult to analyze. Figure XIID again involves a design in which the treatment group is expected to "improve" because of the program and approach conditions found in the comparison group. Regression effects may be a problem in each of the above four designs. Figure XIIE represents a somewhat unusual expectation which, if it takes places, results in a quite convincing design; the treatment group, chosen because of some extreme circumstance, is expected not only to equal but surpass the unchanging comparison group.

Examples of these various outcome patterns may also be identified in Figure XIII which depicts rice crop yields for Pakistan (the program group in our previous examples) and several other countries (comparison groups.) Yields are recorded for the 1966/67 and 1968/69 crop years, the latter being the first in which the impact of high-yield varietals might be expected to show up for Pakistan. (By this time roughly 3-5 percent of rice production area would have been in high-yield varietals.) As noted

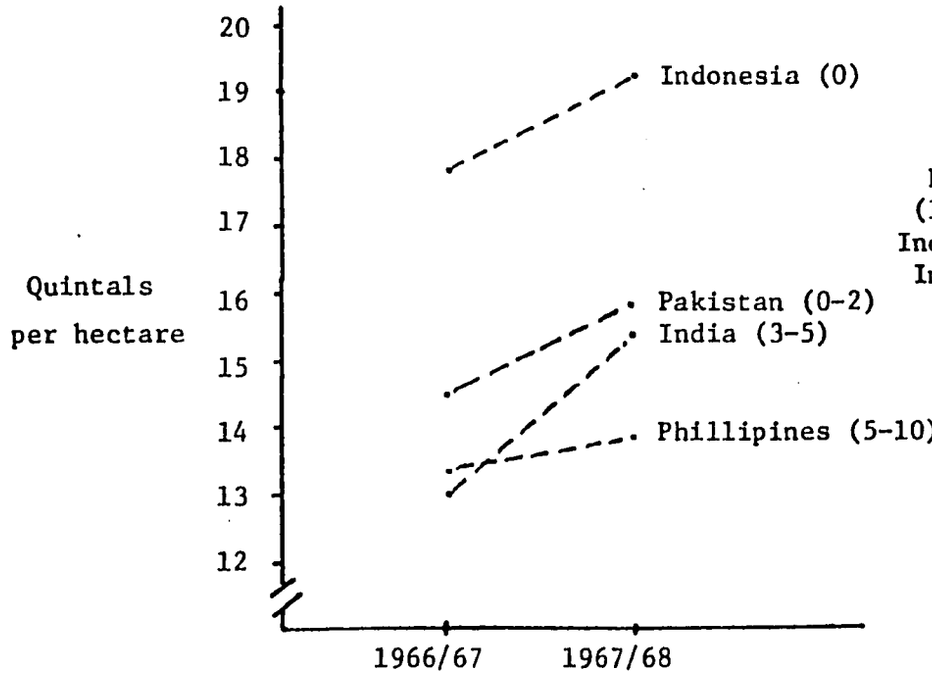


Figure XIII

Rice Crops Yields (1966/67, 1967/68): Indonesia, Pakistan, India, Phillipines

previously, yields in Pakistan did rise approximately 10 percent (14.5 to 16.0). This increase almost exactly paralleled that found in Indonesia, which incidentally had not yet adopted high yield varietals, but was less than the increase shown for India, which had about 5 per cent of its rice land in high-yield grains. Thus, the significance that the evaluator would attribute to the Pakistan results would depend heavily upon which country or countries were chosen as comparison groups. In contrast to Indonesia and India, for instance, if the Phillipines were chosen as a comparison group, the increase in Pakistan would appear quite notable (especially since by 1967/68 in the Phillipines about 20 per cent of the rice land was planted to high-yield varietals.) The reader might also note that the outcomes for India and the Phillipines follow the pattern illustrated in Figure XIIE. One is reluctant, however, to draw any definite conclusions from Figure XIII, since as was illustrated in the hypothetical case of Figure XI, it is risky to make long-term conclusions on the basis of single pre- and post-program measurements.*

---

*Agricultural researchers certainly would not be drawn into such a trap. We are presenting the results in this fashion for illustrative purposes.

While the pretest, posttest design may be a model for the foreign assistance evaluator, often he cannot implement this design because of the absence of adequate data on the pretest or for the comparison group. Two possible variants of the basic design may help to remedy such a situation. First, if pre-program measures are not available, it might be possible through archives to obtain some data as to pre-program conditions, although not necessarily on the same variable as the posttest measure. This could lead to a design like that of Figure XIVA in which the pre-program observations are of a different measure (A, in the diagram) than the post-program observations (B, in the diagram). The possibility of interpreting such a

O_A	X	O_B
O_A		O_B

Figure XIVA

Pretest Posttest Design
Utilizing Different Measures, Example One

design obviously depends on the degree of correspondence between the pre and post-program measures. Relatively complex statistical manipulations are necessary to determine if changes in the correlations between the two are significant; but these may be worthwhile to achieve a better evaluation.[15]

Similar techniques may be utilized in the circumstances indicated in Figure XIVB. Here within group measures are the same, but the comparison

O_A	X	O_A
O_B		O_B

Figure XIVB

Pretest Posttest Design
Utilizing Different Measures, Example Two

group measures (the B's in the diagram) different indicator than the treatment group measures (the A's in the diagram). Again, the soundness of the evaluation depends on the extent of correspondence between the two measures. If dealing, for example, with two different countries, each of which records its average annual income figures in a different manner, the foreign assistance evaluator using this design variant could obtain some conclusions about the country-level effects of large scale programs.[16]

(5) The Interrupted Time Series Design. As shown in Figure XV, this design involves taking a series of observations of the program group before and after the administration of the program. If the effects of the program

0	0	0	0	0	X	0	0	0	0	0

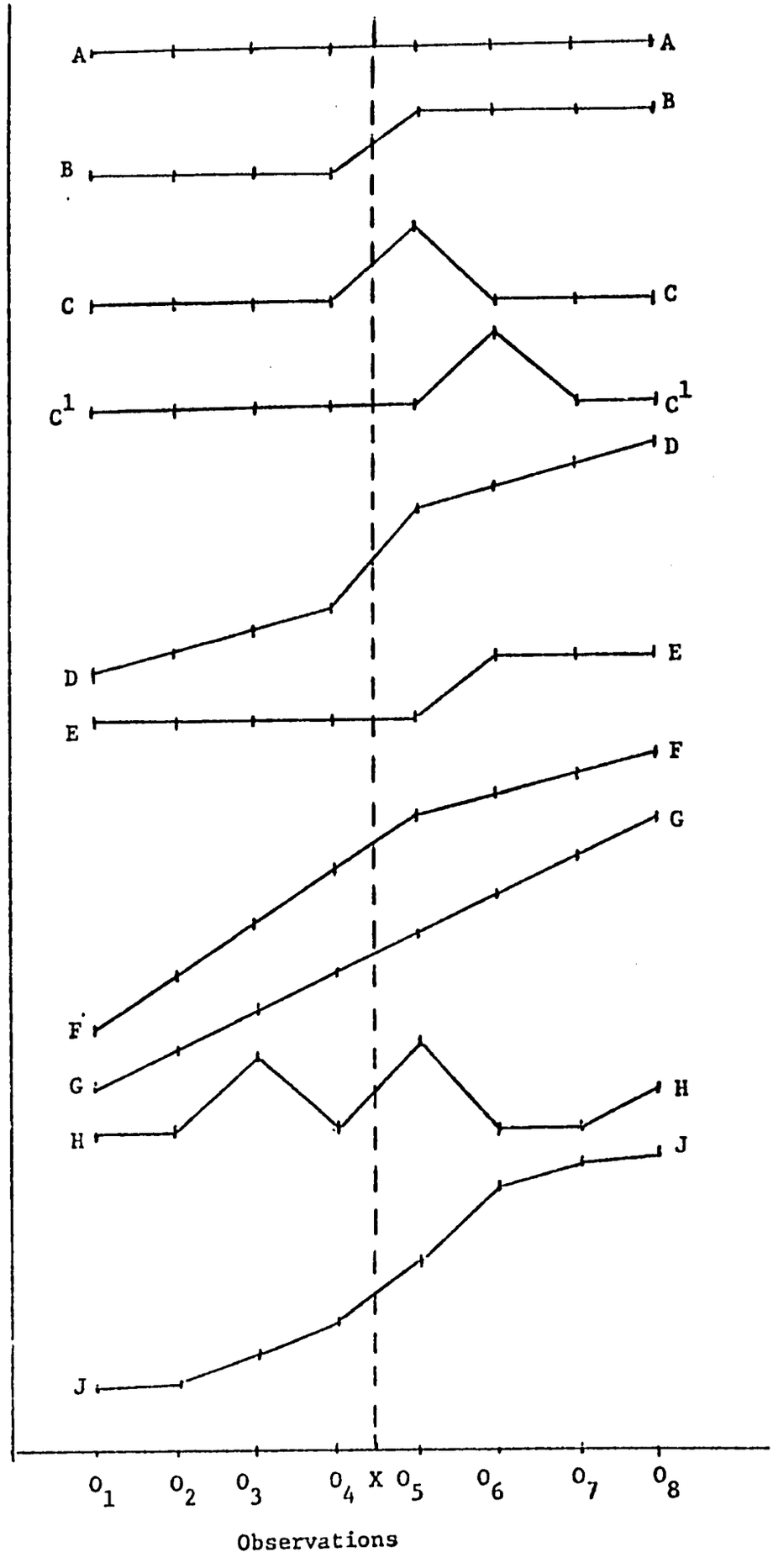
Figure XV
The Interrupted Time Series Design

or project intervention are significant, they will be indicated by a distinct and lasting alteration in the time series pattern. This evaluation procedure therefore guards against the problem of a single fluctuation being mistaken for a significant and lasting treatment effect. The interrupted time series design is a popular one; it is easy to implement and lends itself to graphic display and analysis. Commonly used in the evaluation of domestic social action programs, it is a format that could be adopted often by foreign assistance officials in order to demonstrate program or project effects.

Figure XVI displays a variety of the patterns of possible outcome of interrupted time series design. Lines A, G, H, and J indicate treatments without effect, even though in the latter three cases there is a significant differences between each observation. Lines B and D result from programs

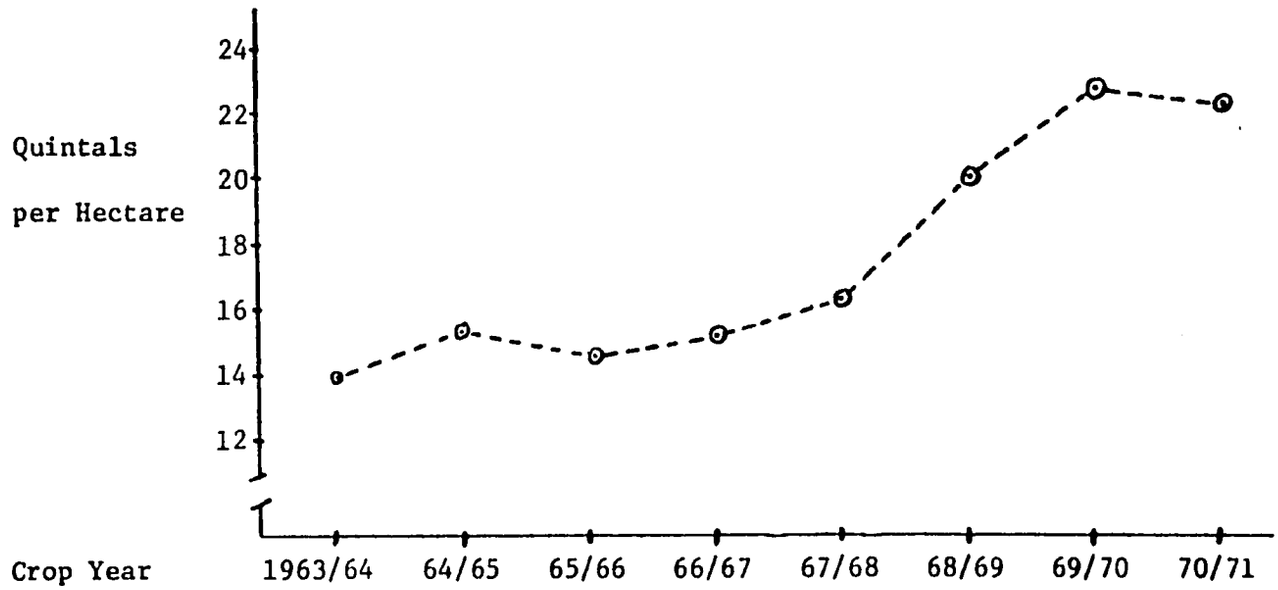
Figure XVI

Possible Outcome Patterns in the Interrupted Time Series Design



which did not alter the basic direction of the pattern (the slope) but altered its level (the intercept, in statistical terms), thus indicating a definite effect. Outcome F, on the other hand, denotes that the treatment brought about a change both in the direction and the level of outputs. All of these results are easily analyzed. Outcomes C, C^1, and E, however, are not because the treatment effect in these situations appears to be a delayed one. Thus, before asserting a significant outcome, the investigator must eliminate the possibility that some extraneous post-treatment event has intervened and caused the observed changes.

When annual rice production figures for Pakistan are displayed in a time series fashion as in Figure XVII, the impact of the new high-yield varietals becomes much easier to evaluate.



Proportion of Area Devoted HYV Rice	0	0	0	0	3	20	30	36

Figure XVII

Annual Rice Production Yields in Pakistan

The observer is given a clear picture of previous trends and ample indication that the supposed impact of the new varietals was not really the result of a single years unusual conditions. Indeed production is seen to grow quite consistently as a greater proportion of rice land is devoted to the HYV's. The advantages of the time-series design as opposed to the single pretest, posttest designs discussed earlier are apparent.

Even given its many advantages, the interrupted time series design is susceptible to a number of threats to validity. Because all of the observations deal with a single group, the effects of maturation or the incidence of history might be mistaken for significant treatment effects. Instrumentation, the "new police chief effect," can also be a serious problem, especially in foreign assistance projects where a relevant data series seldom exists prior to, or remains unaltered by, the implementation of some new program. Finally, quasi-experimental designs of this nature are best suited for identifying relatively abrupt changes in outcome patterns. In situations where the treatment effect is delayed or distributed over time, or where the treatment itself cannot be administered as a one-shot effort, then analysis in this framework becomes exceedingly difficult.

As a concluding note, three "rules of implementation" cited by Reicken and Boruch appear worth remembering about this design:[17]

(a) keep the measurement system constant;

(b) introduce the new program or policy abruptly;

(c) delay reaction to acute problems to avoid statistical regression problems. In other words, do not introduce a new program immediately following some extreme fluctuation in the time series.

(6) The Interrupted Time Series Design with Comparison Group(s). This design, illustrated in Figure XVIII consists of two or more time series of observations, one involving the treatment group, the other(s) involving comparison group or groups. If this format is properly implemented,

O	O	O	O	X	O	O	O	O
O	O	O	O		O	O	O	O

Figure XVII

Interrupted Time Series Design with Control Group Series

i.e. if the comparison group(s) are similar in essential ways to the treatment group and if problems of instrumentation are avoided--then this design is without question the soundest of the quasi-experimental designs. Effects due to history and maturation can be distinguished and to some extent discounted because of the existence of comparison time series. External validity is assured to the extent that the treatment and comparison groups are representative of other environments in which similar programs may be undertaken. Furthermore, a variety of highly sophisticated statistical techniques have been developed which enable the trained investigator to discount the effects of cycles, trends, and seasonal variations in time series, thus allowing an accurate assessment of relatively complicated program effects.[18]

The value of utilizing a control group series is indicated in Figure XIX which illustrates data from the continuing Pakistan rice yields example. Note that the small vertical lines interrupting each of the four national time series indicate the approximate dates upon which the effects, if any, of high-yield varietals should begin to become evident for that particular nation. The Pakistani results are simply repeated from Figure XVII. However,

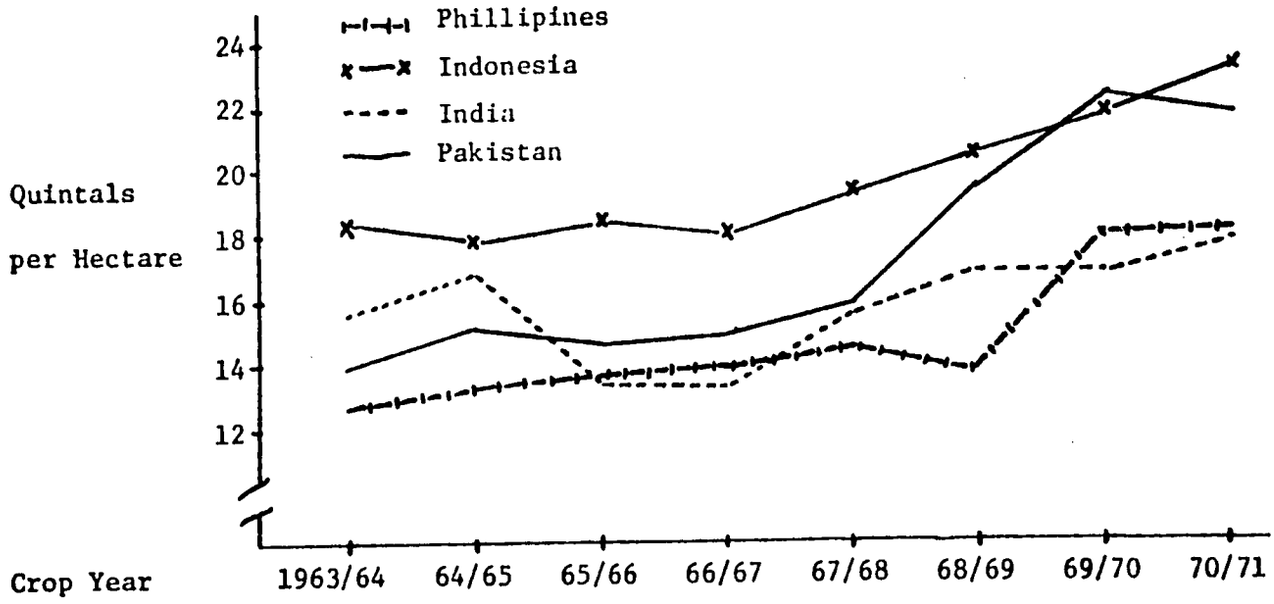


Figure XIX
Annual Rice Production Yields in Four Asian States

the addition of Indonesian, Indian, and Phillipine data provide the evaluator with a much broader basis from which to draw conclusions. Yields in Pakistan, for instance, appear to have increased much more markedly than those in the other three states. Indonesia's rice yields, consistently higher than Pakistan's during the early 1960's, were equalled for at least one year before the end of that decade. Indian production does not seem to have increased as suddenly or as greatly as did that of Pakistan. Responses to high-yield varietals in the Phillipines do not appear as notable as in any of the other nations. All of these rather impressionistic conclusions could be (and have been) checked much more carefully by agricultural experts. The Figure, however, serves for our purposes to illustrate the considerable improvement of this style of quasi-experimental design over the previous five designs we have discussed.

D. Quasi-Experimental Designs: Some Summary Remarks

Perhaps the most useful way of summarizing all of the designs which were discussed above, is by means of a diagram like Figure XX. It indicates the distinctions between each quasi-experimental design concerning (a) the availability of pre-program observations and (b) the availability of other observations for comparison purposes. There are essentially two basic messages for the evaluator in the quasi-experimental design literature. The first is that an attempt should be made to include an equivalent control group in the program design; the more nearly the characteristics of the comparison group match those of the designated treatment group, the fewer threats to validity will occur. The second message is that every attempt should be made to include pre-treatment measurement in the program design. Such a practice makes good sense for two reasons. If the location of a program or project is not fixed, a survey of possible sites will provide useful information about the most suitable locale. And, if pre-treatment measures are utilized in an evaluation, there is a basis for accurate assessment of within-group changes brought about by the program.

NOTES

[1]Thomas Cook and Donald Campbell, "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings," in M.D. Dunnette, ed., Handbook of Industrial and Organizational Research, Chicago: Rand McNally, 1975, p. 2.

[2]Ibid.

[3]Henry Reicken and Robert Boruch, eds., Social Experimentation: A Method for Planning and Evaluating Social Intervention, New York: Academic Press, 1975, p. 2.

[4]Figure IV is an adaptation of a similar illustration in Carol B. Weiss, Evaluation Research, Englewood Cliffs: Prentice-Hall, Inc., 1972, p. 61.

[5]For a forceful presentation of these views see: Edward A. Suchman, "Action for What? A Critique of Evaluative Research," in Carol B. Weiss, ed., Evaluating Social Action Programs: Readings in Social Action and Evaluation, Boston: Allyn and Bacon, p. 64.

[6]Henry Reicken and Robert Boruch, op.cit.; Robert Boruch and Henry Reicken, Applications of Randomized Experiments to Planning and Evaluating AID Programs," Final Report of AID Grant no. AID/cm/ta-c-1055, October, 1974, pp. 30-33. See also Ilene Bernstein and H.E. Freeman, Academic and Extrepreneurial Research: The Consequences of Diversity in Federal Evaluation Studies, New York: Russell Sage, 1975.

[7]I. Bernstein and H. Freeman, op.cit.

[8]See especially Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research, Chicago: Rand-McNally, 1966; and Thomas Cook and Donald Campbell, op.cit. The latter contains an extensive bibliography in which numerous of Campbell's other works are cited.

[9]Henry Reicken and Robert Boruch, op.cit., p. 8.

[10]For further description of this design see: Ibid., pp. 115-116, and Thomas Cook and Donald Campbell, op.cit., pp. 80-82.

11

[12]Thomas Cook and Donald Campbell, op.cit., p. 68.

[13]Robert Boruch and Henry Reicken, op.cit., cite several examples of foreign assistance project designs involving cohort groups.

[14]Thomas Cook and Donald Campbell, op.cit., pp. 38-48.

[15] Ibid., p. 49.

[16] For another example of this type of design see ibid., pp. 61-64.

[17] Henry Reicken and Robert Boruch, op.cit., p. 107.

[18] A brief introduction to questions of time series analyses in quasi-experimental designs is provided by James Caparaso in his "Quasi-Experimental Approaches to Social Science," in James Caparaso, ed., Quasi-Experimental Approaches: Testing Theory and Evaluating Policy, Evanston, Ill.: Northwestern University Press, 1973, pp. 3-38.

For more complex and fascinating examples of time series analyses see: Donald Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time series Data in Quasi-Experimental Analysis," Law and Society Review, vol. III, no. 1, 1968, pp. 33-53; G.V. Glass, G.C. Tiao, and T.O. Maguire, "The 1900 Revision of German Divorce Laws: Analysis of Data as a Time-Series Experiment," Law and Society Review, vol. 6, 1971, pp. 539-562; H.L. Ross, D.T. Campbell, G.V. Glass, "Determining the Effects of a Legal Reform: The British Breathalyser Crackdown of 1967," American Behavioral Scientist, vol. 13, 1970, pp. 493-509; G.V. Glass, V.L. Wilson, and J.M. Gottman, Design and Analysis of Time Series Experiments, Research Report, Laboratory of Educational Research, University of Colorado, 1972.

Chapter IV

Design and Evaluation of Foreign Assistance Activities

There are no pat answers to the decision maker's questions about which design is the "best" evaluation format. Circumstances, personnel, and finances differ from one project to another so considerably that what will work in one situation may fail in the next. The message in our previous discussions, however, has been that the decision maker should adopt the soundest evaluation design possible. Often with a minimal amount of effort or expense, a weak design may be altered to produce a much stronger one. With this in mind, our attention in this final chapter is devoted to an assessment of the utility of experimental and quasi-experimental design principles for AID evaluation activities. Consideration of project-level evaluations will be separate from a discussion of program, sector, and country-level evaluations.

A. Sounder Design of Project-Level Evaluations

AID devotes considerable time and resources to project evaluation. Each individual project is supposedly reviewed annually while it is ongoing, with a complete study being done when the project is concluded. Information gained from year to year and from project to project should provide the decision maker with knowledge about which types of projects have the most impact, which techniques are the most cost-effective, etc. Thus, while the circumstances and specific goals of each individual project evaluation may differ, the basic goals of proving whether or not assumed causal relationships exist and of determining whether or not results are significant remain the same.

Upon reviewing the catalogue of threats to validity and the various experimental and quasi-experimental procedures that can be adopted to circumvent these threats, three central aspects of sound evaluation design are apparent. (1) Empirical information and data should be collected prior to the implementation of a project. In other words, evaluations should involve pretests. (2) Wherever possible project designs should include control groups or comparison groups so that evaluators have information bases from which to draw comparisons. (3) The recording of observations beginning before the project delivery starts and lasting until well after all foreign and official personnel have been recalled are of critical importance if the evaluator is to draw conclusions about trends, long-term effects, and unanticipated consequences. In other words, time-series data gathered on treatment and control or comparison groups are an important asset to the evaluator. Further explanation and comment on each of these points may prove useful.

1. The necessity of pretests.

Meaningful evaluation on the basis of posttest data alone, i.e. data gathered after the project is ongoing, is almost impossible. (Recall design 1 in the previous chapter.) The logic of evaluation is comparative in nature; without some reference point or points as to conditions prior to the intervention there is no way to assess the extent of changes induced by the inputs. In unusual circumstances, such as disaster relief projects, advance monitoring of the target population may not be possible. "Evaluation" of such projects is restricted to reporting and describing the extent and the schedule of the assistance. (One proceeds on the assumption that food, clothing, shelter, and medicine are effective remedies, the important issues become those of volume and extent of coverage.)

Very few, if any, AID projects are begun without some attention to the matter of pre-project measurement. "Beginning-of-project-status" (BOPS in the log-frame terminology) is to be assessed prior to the commencement of each project.[1] This does not mean, however, that pre-project data are always recorded on indiciators which are useful or relevant to post-project evaluation. Therefore, it must be stressed that in order to implement sound evaluation designs comparable data on measures relevant to the project activity, goal, and purpose must be utilized. This implies careful advance planning by both project personnel and evaluators.

Special attention should be given to those designs which utilize post-treatment measures only but which involve control or comparison groups. These procedures, as discussed in conjunction with design 2 in the previous chapter, entail collecting information on groups of individuals that are presumably at a similar stage of maturation, education, or skill as would be expected of the target group individuals had they not participated in the peoject. Selection effects and testing effects may threaten the validity of these designs. Furthermore, great caution should be exercised in choosing cohort groups to ensure their equivalency on all essential characteristics. As pointed out before, variants of the posttest only design, if utilized with care, could prove very useful in some situations encountered by AID evaluators.

2. The importance of control and/or comparison groups.

With data on the treatment group before, during, and after the delivery of project inputs, the evaluator is able to report changes over time within this particular subgroup. However on this basis alone he can not make judgements about the success or effectiveness of the project since he can not discount threats to internal and external validity. That is, he has no way of knowing if the effects observed were confined to the treatment group or if the treatment group itself was representative of any broader population.

These assurances can only be provided by including control or comparison groups in the evaluation design. It is difficult to overstress the central importance of comparison or control groups--without them between-group comparisons are impossible, without them an evaluator is severely limited in drawing any conclusions about a project.

The soundest method of choosing and assigning individuals (persons, families, villages, etc.) to separate groups in a project is through randomization. As indicated in Chapter III, groups selected in this manner are termed control groups, and designs involving them are experimental designs. In situations where groups are chosen by nonrandom procedures, we label these groups comparison groups instead, and such designs, quasi-experimental designs. Experimental designs provide sounder bases for evaluative judgements than their quasi-experimental counterparts because random assignment and selection are the only sure ways to achieve equivalency among groups.

This does not mean, however, that experimental designs have been, or will be, utilized extensively in foreign assistance project evaluations. Decision makers offer a variety of explanations for this fact.[2] (a) Cost and resources: the setting up and administration of a well-controlled field experiment is expensive, time-consuming, and requires trained personnel. Often these may simply not be available to AID. (b) Political opposition: Administrators, local officials, and participants often oppose experimental projects because they do not wish to see individuals assigned by "chance" or denied treatment in case of need. Even though, as we argued earlier, most of these arguments imply a misunderstanding of good field experimental procedures, if opposition to randomization is so strong as to jeopardize the project as a whole, then alternative measures may have to be adopted.

(c) Situational conditions: In certain project environments random selection of control groups may not be feasible. For example, if an irrigation project is to be established in a river valley, it may be impossible not to provide all farmers or villages with the same amount of water. A comparison group of farmers in another valley would have to be selected for evaluation purposes.

(d) Project purposes: Experimental designs are most appropriate for, in fact are the only sound format for, assessment of the relative effects of different input schedules, on different treatments, within a project. If the evaluator wishes to find out, for example, which of a series of educational and medical regimens is most effective in achieving the adoption of fertility-control methods by rural women, he should utilize an experimental design for this purpose. Similarly, if he wishes to assess the effectiveness of different TV education programs, he should monitor a series of treatment and control groups. However, in some circumstances AID officials are not interested in testing alternatives procedures within a project. They may feel assured that the method involved is appropriate and has been proven on other occasions. For example, projects designed to introduce high yield grain varietals to developing countries are no longer usually designed to assess the utility of the new types of grain. This issue has been settled previously in extensive agricultural research. What decision makers wish to evaluate now is the overall impact of such projects on local grain production, and they are not apt to see the utility in conducting an experimental design to do this.

Whatever the reason, in those situations where an experimental design is not going to be used the decision maker should make every effort to employ the soundest possible quasi-experimental design. This will usually mean the use of comparison groups in lieu of control groups; and if chosen with care, sound

evaluation will certainly be possible. It should be emphasized that we believe that experimental procedures ought to be given first consideration in the design and evaluation of any project. If and when it is found that circumstances prevent pre-project measurement, or random selection and assignment of control groups, then the administrator or evaluator may wish to revert to one of the series of quasi-experimental designs discussed previously. But this decision to make do should come as a result of practical difficulties--not direct planning from the pre-project stage forward. Quasi-experimental techniques, admittedly, acquire much of their popularity and utility because they are flexible and adaptable to adverse conditions. Too often, however, one gets the impression that administrators might be willing to approve projects with less-than-optimal characteristics, knowing in advance that at least some variety of quasi-experimental evaluation might work--thus satisfying organizational requirements.

In their 1974 report to AID concerning the agency's evaluation procedures, Robert Boruch and Henry Reicken advocate increased utilization of experimental designs.[3] They argue that such procedures are particularly suited for certain situations which AID encounters in many of its project environments. Specifically, they note that in education projects and in training projects where permanent facilities exist and regular, formal classes are conducted experimental evaluations are relatively easy to construct. Fertility control projects and other socio-medical, nutritional projects also lend themselves to an experimental format as do some types of communication and agricultural projects. By citing numerous examples of domestic and foreign evaluations Boruch and Reicken compile a strong case.

The two authors also help to clear away some of the misconceptions surrounding experimental and quasi-experimental procedures. They highlight the fact that control groups do not have to be deprived of treatment in order

to be evaluated and that experimental and quasi-experimental methods may be artfully combined in a single project. Brief descriptions of three actual foreign assistance projects (two of which are described in the Boruch and Reicken report) will serve to illustrate these points.

(i) Cali, Columbia Project referred to under full title as "Stimulation of intellectual and social competence in Columbian preschool children affected by multiple deprivations of depressed urban environments." This project conducted by the Universidad del Valle and funded mainly by the Ford Foundation is intended to assess the effects of nutritional, educational, medical treatments separately (and in combination) in alleviating social and intellectual deprivation of children in depressed urban environments.[4]

To implement the project, a series of appropriate neighborhoods were identified; these neighborhoods were then randomly assigned to different treatment groups in which treatment was administered to all nutritionally deprived three-year old children in the population. Each year a "lottery" was held to determine the participants in the various treatment groups for that year. Some children were in groups which received medical, nutritional, and educational treatment. Others received just medical and nutritional treatment, while those in control groups received medical attention if social workers noted it was needed. A group of non-deprived (middle class) children were observed to provide a separate, quasi-experimental control group for the project. Pretest and posttest measurements were administered.

This design therefore has several interesting features as sketched in Figure XX. Random select and assignment was not done on individuals per se, but rather on small neighborhood communities of individuals. Children that interacted together thus were in the same treatment group. Initial selection of the neighborhoods was done on the basis of specified economic and social characteristics. The groups themselves received varying degrees of treatment

Figure XX

Outline of the Design of the Cali Columbia Project

Selection of Qualified Urban Poor Neighborhoods

↓

Random Assignment of Neighborhoods to Treatment Groups

Treatment	A Groups	B Groups	Control Group	Quasi-Experimental Control Group
Educational Treatment	✓			
Nutritional Treatment	✓	✓		
Medical Treatment	✓	✓	Only as Needed	

to facilitate assessment of the relative effectiveness of nutritional, medical, and educational approaches. The involvement of the group of middle class youngsters as an "external" control is a particularly interesting design feature.

(ii) Taichung, Taiwan Family Planning Project. The city of Taichung was chosen as the site for an intensive family planning project--a precursor to a subsequent national family planning program. Work was carried out in 1963 and 1964 by the Taiwan Population Studies Center and the University of Michigan Population Center.[5] The aims of the Taichung project were to discover which types and combinations of treatments were most effective in getting couples to adopt family planning and to discover which method of family planning was most favored by the population. As in the Cali Columbia Project the population was divided into small groups, these small groups then being randomly assigned to treatment groups.

The neighbor unit utilized in the Taichung project was called a lin, each lin composed of about 20 households and about 12 married women aged 20-39. The city of Taichung contained about 2400 such units and approximately 36,000 married couples in the 20-39 age bracket. Random assignment placed each lin in one of four treatment groups: (1) a so-called "everything--husband and wife group" (Ehw) in which both husband and wife were personally visited, mailings were sent, and neighborhood meetings were held, (2) an "everything--wife only group" (Ew) which involved all of the treatment given to the Ehw group except the visit to the husband, (3) a "mailings group" (m) which received only mailings, and (4) a control group in which couples were not contacted directly at all. Further separations were made according to the density of the districts. Pre and post-project interviews were conducted along with medical surveys.

The structure of this design was essentially experimental, with the minor exception of the random assignment of lins rather than couples. Figure XXI provides a summary of the Taichung design.

Figure XXI

Outline of the Taichung, Taiwan Family Planning Project

Demarcation of City into 2400 Lins

↓

Random Assignment of Lins to Groups

Treatment	Ehw Group	Ew Group	M Group	Control Group
Visit: Husband	✓			
Visit: Wife	✓	✓		
Mailings	✓	✓	✓	

The results of the Taichung Project indicated that some 80% of those couples (5424) which accepted family planned chose the IUD. Mailings by themselves were not effective nor did husband visitation prove to be worth the additional trouble. Thus in the national family planning program commenced shortly afterwards, emphasis was placed upon adoption of the IUD with an "Everything--wife only" schedule treatments. This example illustrates how an experimentally designed project may provide a valuable and efficient testing ground of methods and techniques to be used in larger programs.

(iii) Danfa, Ghana Family Planning and Rural Health Project: The Danfa project is a very large multi-purpose training, research, and institutional development project undertaken by the Ghana Medical School, UCLA advisors and personnel, with AID funding.[6] Initial planning was underway in 1970; work still continues on this project. One of the goals of the project was to examine the effectiveness of family planning services when delivered in conjunction with other health services in the rural Ghanian environment, and it is these aspects of the project design which are of interest to us.

In order to test the effectiveness of the various treatment programs, four separate rural areas north and west of Accra were selected. Chosen to be similar on all relevant health demographic, and socio-economic indicators, each area contained 10,000 to 15,000 people. Individuals in area I received family planning services as well as health education and comprehensive health care. Those in area II received family planning and health education services; while those in area III received only family planning services. Area IV served as a "control" area in which individuals received "ordinary" Ministry of Health (MOH) services common to all other regions of the nation. The basic design was therefore quasi-experimental in that the treatment units were not chosen randomly. The four areas constituted nonequivalent control groups instead, as indicated in the Figure below.

Figure XXII

Outline of the Danfa Family Planning Project

Services Delivered	Area I	Area II	Area III	Area IV
Health Care	Comprehensive	MOH	MOH	MOH
Family Planning	✔	✔	✔	
Health Education	✔	✔		

Random Selection of 500 households per group for Baseline Longitudinal Survey

Random Selection of 20 villages per group for Village Health Survey

While the basic framework of the Danfa project design was quasi-experimental, most of the measurement and data-collecting procedures, on the other hand, were based on random selection techniques. Thus, for instance a series of surveys to establish longitudinal baseline data on morbidity, fertility, and health practices were administered to some 500 households chosen from each of the four areas. In order to closely monitor the incidence of illness, immunization, and sanitation programs a Village Health Survey was set up. Based on a random cluster sampling of 20 villages (5 per each area), a survey of some 1000 individuals was conducted in 1972-73 with second and third repeat interviews scheduled for 1975 and 1977.

Thus the Danfa project illustrates a different combination of experimental and quasi-experimental procedures from those found in the Cali and Taichung projects. In the latter, two small population units were randomly selected and assigned to groups; in the Danfa project large population aggregates were selected first and then individuals were randomly selected from within these areas for testing purposes. Each of the above examples demonstrates not only the importance of including groups within a

project design for comparison purposes, but also some of the flexibility and variety of options available to project design and evaluation officers.

3. The value of time-series information.

One of the most difficult tasks in evaluation is the assessment of long-term project effects. Studies which involve only one-shot pre- and post-project measurements, or studies which involve a series of measurements only during a project's active life can not accurately account for long term trends or delayed, unexpected consequences of the treatments. This can only be done by recording and analyzing a series of observations, beginning before the project starts and running until well after the administration of treatment has been completed. Thus, the final important principle of evaluation is to include, whenever possible, time series observations in the design. The major flaw in the otherwise strong pre-test, post-test design involving control or comparison groups (design 4 in the previous Chapter) is that it is highly susceptible to short-term variations. The decision maker cannot be assured that pre-project observations were sharply at variance with previous trends; nor can he be confident that in two or three years from now the projects impact will still be apparent. Time series designs, especially those involving comparison groups can avoid these uncertainties and errors. Designs 5 and 6, the time series designs, thus are stronger and sounder formulations than the other design types outlined in Chapter II.

The major practical drawback to the time series designs is that they require pre-project measurements on indicators which may be sensitive to project effects. Most often appropriate data does not exist in advance; furthermore, project personnel will not have the time or finances to conduct several separate measurements prior to initiation of the delivery of actual project inputs. Often if time series data does exist, the information will

likely be on measures which are irrelevent or insensitive to the project effects. Thus, for example, while estimates of average birthrate for the nation may be available in national archives, these figures are probably not going to be of great utility to someone who has to assess the effects of a foreign assistance project disseminating family planning services in several select small rural villages. Only occasionally will a project be large enough that its effects may have such impact that social archive aggregate data will be useful for time series evaluation studies. On the other hand, as Boruch and Reicken point out, certain types of AID project situations can be more readily evaluated in a time series fashion than others. These include educational projects where students are likely to have been tested annually since they began school, or agricultural projects where some annual crop production statistics have been previously collected.[7]

B. Program Level, Sector Level, and Country Level Evaluations

Project level evaluation is only one phase of the evaluation task facing most government agencies such as the A.I.D. Program level, sector level, and country level evaluations must be undertaken as well. However, evaluation at these levels is done in different fashions and for different reasons than project level evaluations. More specifically, there are two distinctive aspects to such investigations: the time span involved in program, sector, or country level studies probably will be longer than for a project-oriented investigation, and the focus of program, sector, or country level studies generally will be broader than that of a project evaluation. In the latter situation, the investigator will most likely be concerned with specific questions, e.g. which method of treatment was most efficient, which type of individual unit (person, farm, or village) was most receptive, was the project

an efficient one, etc.? In a program evaluation, on the other hand, attention will be given to larger issues, e.g. which project, or even which type of project, was most successful; what was the over-all impact of the program upon the province, region, or country in which it was implemented; how does this program compare to those in other countries, etc.? In other words, in program, sector and country level evaluations, interest is concentrated more on demonstrating whether or not the agency's efforts have had an effect, any effect, than upon testing whether or not a particular treatment or project within the larger plan of action was efficient or successful.

This shifting of concern and level has implications for the evaluator. The designs and techniques appropriate for project evaluation will no longer be practicable. Program, sector, and country level studies require planning, resources, data, and personnel of a quite different nature and usually for greater magnitude. Federal domestic program evaluation, for example, is a complex field of study in itself. However, while numerous volumes have been written on domestic evaluations, very little has been done on large scale evaluation in the foreign assistance field. It is beyond the scope of this monograph to do so as well; but in the context of the previous experimental design--quasi-experimental design exposition, there are three issues which should be noted.

First of all, designs that entail only one-shot pre and post-program measurements or evaluations that, eventhough they involve numerous measurements, are completed in a relatively short period are less appropriate for "higher level" studies. If a fertility control program, for example, took five years to implement its effects would be only partially evident at the end of this period. Perhaps another five to ten years would have to pass before any final assessment could be made. Thus it would be naive to take a single observation

before the administration of the program began and attempt to link it to a single post-program measure, taken at the immediate end of the treatment, and regard such a design as a meaningful evaluation. Not only would it be impossible to discern the project's full impact, but with only two measures widely separated in time, threats of validity due to history, maturation, selection, and mortality would be hard to discount. Time series designs, utilizing control or comparison groups wherever possible, thus are virtually essential if a major program or sector or multi-country endeavor is to be accurately monitored and evaluated.

The second point concerns the type of data employed in large scale evaluations. Project data will usually consist of individual data, which, if it is aggregated, will be for small units, such as production or income figures for farmers in a specified village, birthrates for small village units, test scores for classes of students, etc. However, program, sector, and country-level evaluations are more likely to require aggregate data, that is data that concerns larger units and possibly longer time intervals. Thus, for example, sector or country studies will utilize measures such as average annual national birthrates, average annual small farmer income, annual crop production figures, etc. This means that it is unlikely that the evaluator or even the agency performing the study will be able to collect any or all of the relevent information by themselves. Pre-program measures will have to be extracted from provincial or national archives. Host country officials will have to be relied upon to continue making observations after foreign advisors and administrators have left the country. The evaluator, thus, may be dependent on others for his information. Not only may the existence of adequate time series data be problemmatical, but the accuracy and comparability of the observations may be suspect. There is no

easy remedy to this basic problem. Evaluations may have to be done despite the quality of the data; the decision maker has to build in margins for error in his calculations and try to verify his findings by utilizing information from other sources. (For example in evaluating agricultural projects, crop yield data may be corraborated by examining farm income data on market volume, data fertilizer and seed purchase data, etc.)

Finally the utilization of control and comparison groups poses a major problem in program, sector, and country-level evaluation. This is really a two-faceted problem: one facet concerns the internal aspects of the program, i.e. whether or not control groups and experimental procedures have been utilized in the administration of the program; the other facet concerns external aspects of the design, i.e. whether or not other programs or nations exist which may be employed as comparison groups.

To construct a full-scale program on an experimental basis is a very expensive and time-consuming endeavor. This is seldom done in the domestic environment--the negative income tax experiment being a notable exception--and almost never done in the foreign assistance environment. AID officials express the view that they do not have the resources to undertake research of this nature, nor is it necessary to do so in many of their programs. For example, in medical programs decision makers assume that their services are useful; they are not interested in testing the relative effectiveness of different techniques, they want an assessment of the overall impact of the program. Thus control or comparison groups do not play a role in their evaluation designs. However, as Boruch has argued, this attitude is sometimes a bit too narrow.[8] While it may not be feasible or necessary to establish an entire program on an experimental basis, certain segments of the program may be conducted experimentally in order to facilitate comparisons within the

program or to provide advance information about possible future modifications to the program. To use his phrases, programs may be designed which involve "experiments within quasi-experiments," "quasi-experiments within experiments," or "natural experiments and controlled experiments in tandem." The role of the Taichung family planning experiment (discussed earlier) as a pilot venture in Taiwan's national family planning program provides a very good example.

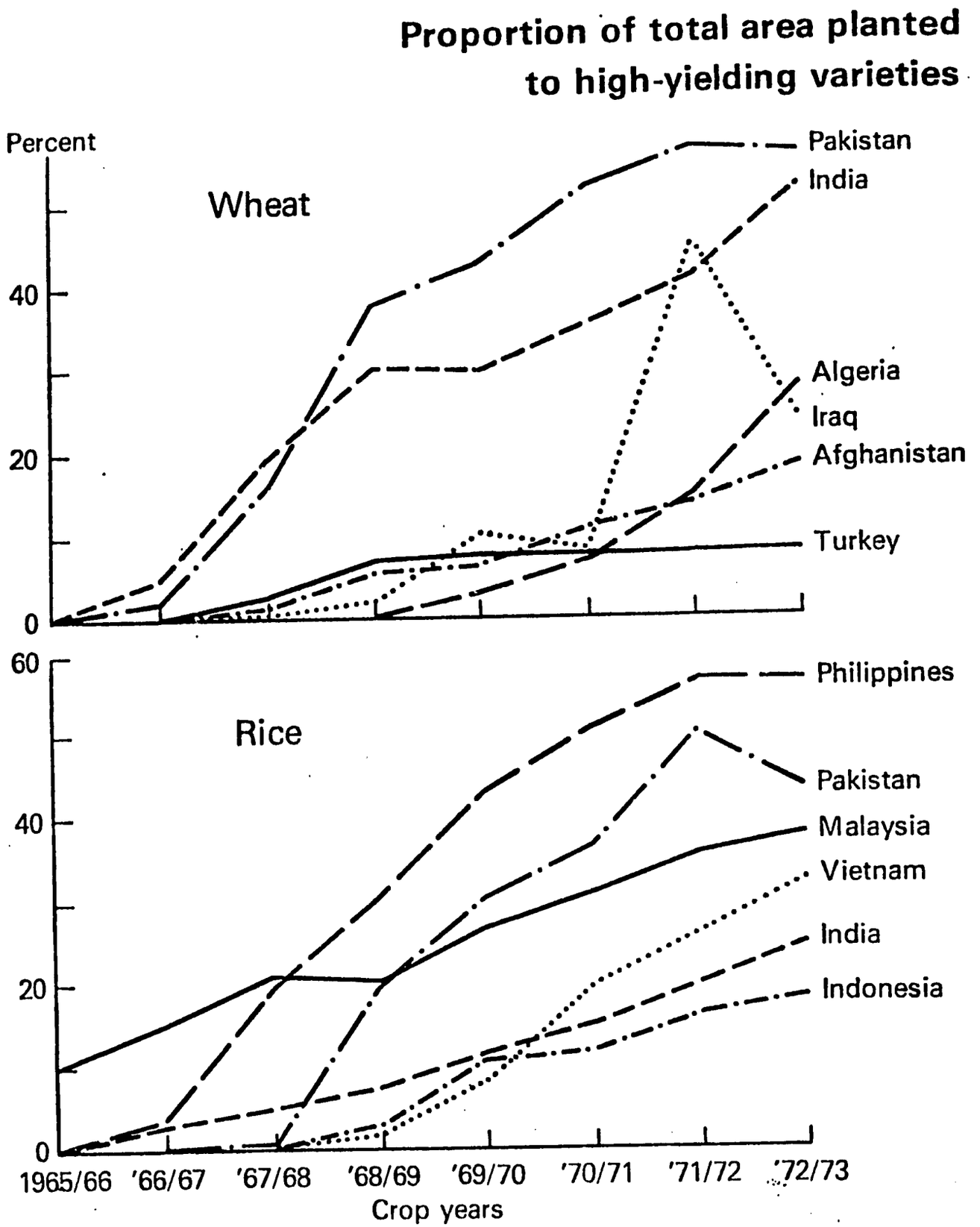
As for the second aspect of control or comparison, random selection of other programs or nations does not make sense in most higher-level evaluations. If evaluating the fertility control program in Taiwan, or assessing the agricultural sector in Peru, the use of the United States or Nigeria or some other selected nation as a control nation will not be fruitful. In such instances, the evaluator must resort to designs that involve only the treatment population, e.g. the interrupted time-series design (no. 5 in the previous chapter), or he must choose comparison groups which are equivalent to the treatment group. Thus, perhaps a country like Taiwan with an intensive birth control program may be compared to another Asian nation which does not have a program of the same variety. A word of caution, however; as one works with larger and larger aggregate units, or units which are further and further separated in time and space the possibility of accounting for extraneous alternative explanatory factors becomes less and less. Thus any comparison group, especially another country, should be chosen with extreme care. Quasi-experimental designs can be very effective in program, sector, and country evaluations, but only to the extent that threats to validity are identified and avoided in advance.

Throughout Chapter III a series of illustrations were constructed citing annual Pakistani wheat yields before and after the introduction of high yield varietals. While these figures were genuine, the examples were

out of context; they were abstracted from a larger, evaluation report entitled Measuring the Green Revolution: The Impact of Research on Wheat and Rice Production.[9] As this larger study in itself constitutes a fine example of the use of quasi-experimental design in a multi-country program evaluation, we thought it appropriate to conclude with a brief look at its methods and results.

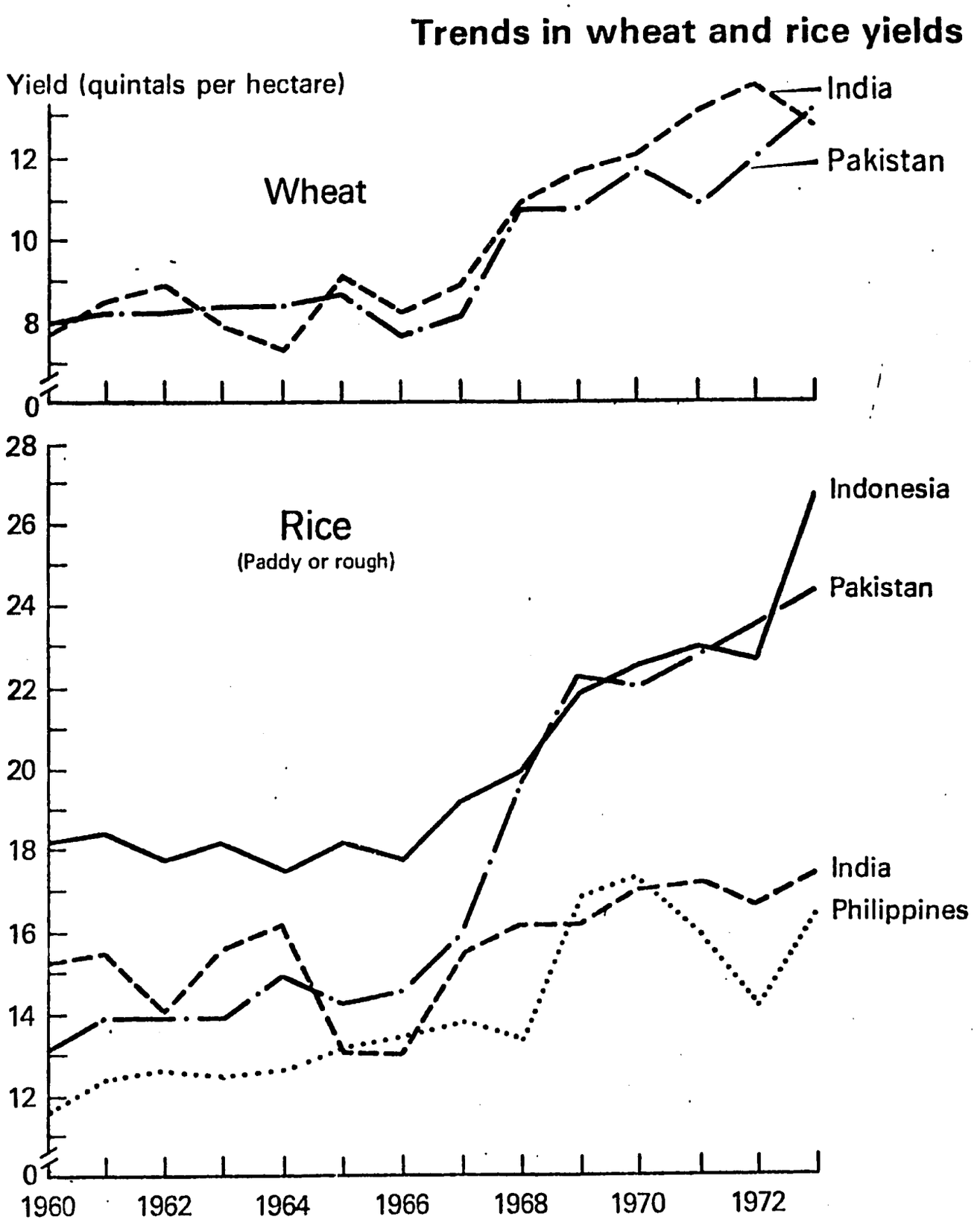
High yield wheat and rice varietals were developed and continue to be developed, in agricultural research programs that commenced about fifteen to twenty years ago. This research was experimental both in conception and design, i.e. the scientists were searching for new varietals and they employed experimental procedures to do so. Around 1965 high varietals of wheat and rice became available for distribution in Asian countries; AID and other development agencies embarking on ambitious programs to encourage their usage. The study cited above was undertaken to assess the impact of the so-called "Green Revolution" in these nations for the 1965/66 - 1972/73 period. Initial national results are reproduced in Figures XXIII and XXIV. They indicate that greater proportions of crop land were planted to high yield varietals and that generally speaking yields rose over the eight year period. Furthermore, Figure XXIV illustrates clearly the format of a time-series design where all of the nations serve as comparison groups for the others. It should be noted that while these diagrams present a relatively clear case, additional sophisticated analysis was necessary to determine the actual effect of the high yield varietals on wheat and rice production in (non-Communist) Asia. These analyses suggested that increases of approximately 28 percent for wheat and 20 percent for rice were realized by 1972/73.[10] While definitely a major contributing factor to this production rise, the high yield varietal program by itself could not account for all of this impact.

Figure XIII



Source: Dana G. Dalrymple, Measuring the Green Revolution: The Impact of Research on Wheat and Rice Production, Economic Research Service, USDA Foreign Agricultural Economic Report No. 106, June, 1975.

Figure XXIV



Source: Dana G. Dalrymple, Measuring the Green Revolution: The Impact of Research on Wheat and Rice Production, Economic Research Service, USDA Foreign Agricultural Economic Report No. 106, June, 1975.

This example, therefore, serves to highlight the importance and utility of sound evaluation procedures, specifically quasi-experimental design methods, for constructing assessments of large scale foreign assistance programs.

NOTES

[1]Refer to the Evaluation Handbook, 2nd ed., Office of Program Evaluation, USAID, Washington, May, 1974; or Project Evaluation Guidelines, 3rd ed., Office of Department Program Review and Evaluation, USAID, Washington, D.C., August, 1974.

[2]For a discussion of these issues see Henry Reicken and Robert Boruch, eds. Social Experimentation: A Method for Planning and Evaluating Social Intervention, New York: Academic Press, 1975.

[3]Robert Boruch and Henry Reicken, "Applications of Randomized Experiments to Planning and Evaluating AID Programs," Final Report of AID Grant no. AID/CM/ta-c-1055, 1974

[4]The information on the Cali, Columbia project is taken from Boruch and Reickens' report, ibid. Original documentation was unavailable to the author.

[5]The research design and results of the Taiwan family planning project are presented in detail in Ronald Freedman and John Y. Takeshita, Family Planning in Taiwan: An Experiment in Social Change, Princeton: Princeton University Press, 1969. See especially pp. 109-148.

[6]The research design of the Danfa project is presented in detail in A.K. Neuman, F.T. Sai, and S.R.A. Dodu, "Danfa Comprehensive Rural Health and Family Planning Project: Ghana, Research Design," Journal of Tropical Pediatrics and Environmental Child Health, February, 1974.

A comprehensive evaluation of the Danfa project was undertaken in 1975 by an outside team of medical and public health experts. Their report is contained in "An Evaluation of the Danfa Comprehensive Rural Health and Family Planning Project in Ghana," prepared by Henry M. Gelfand et al, published by the American Public Health Association, in agreement with USAID (Authorization AID/csd 2604, T.O.17; Ltr POP/FRS: 2/25/75; ALPHA Assgn. No. 146).

[7]Boruch and Reicken, op.cit., pp. 25-30.

[8]Ibid as well as Robert F. Boruch, "Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation," Sociological Methods and Research, vol. 4, no. 1, August 1975, pp. 31-53.

[9]Dana G. Dalrymple, Measuring the Green Revolution: The Impact of Research on Wheat and Rice Production, Economic Research Service, USDA Foreign Agricultural Economic Report No. 106, June, 1975. This is a quite general publication which contains reference to numerous more technical reports on this topic.

[10]Ibid. pp. 29-39.