The EVALUATION Project

Evaluating Family Planning Programs

with Adaptations for Reproductive Health

Bertrand
Magnani
Rutenberg

EVALUATING FAMILY PLANNING PROGRAMS

WITH ADAPTATIONS FOR REPRODUCTIVE HEALTH

Jane T. Bertrand

Robert J. Magnani

Naomi Rutenberg

The EVALUATION Project

September 1996

USAID Contract Number: DPE–3060–C–00–1054–00

Acknowledgments

We wish to acknowledge the support of the Office of Population, United States Agency for International Development (USAID Contract Number: DPE–3060–C–00–1054–00) and thank them for making this publication possible. The views expressed herein are those of the authors and not the sponsoring agency.

The authors wish to thank several members of The EVALUATION Project for their thoughtful critiques of different versions of this manual: Amy Tsui, David Guilkey, Albert Hermalin, John Ross, John Stover, David Hotchkiss, Jennifer Strickler, and Lisanne Brown. Krista Stewart, Technical Advisor to The EVALUATION Project at USAID, also provided useful suggestions on earlier versions.

We also extend our thanks to several external reviewers: Jerry Bowers, John Casterline, Allan Keller, Barbara Janowitz, F.M. Mburu, Virginia Paine, and Jack Reynolds. A number of colleagues provided input on specific aspects of reproductive health (Chapter VII): Gina Dallabetta, Erin Eckert, Susan Hassig, Cate Johnson, Marge Koblinsky, Miriam Labbok, Therese McGinn, Barbara Mensch, Nancy Mock, Leo Morris, and Chloe O'Gara.

Several USAID Missions and country programs have offered The EVALUATION Project staff the opportunity to test different approaches to family planning evaluation at the field level. We are particularly grateful to our colleagues at USAID and the family planning programs in Brazil, Ecuador, Honduras, Ghana, India, Kenya, Morocco, Nigeria, Peru, Philippines, and Tanzania for their support of and participation in a number of evaluation activities from 1992 to the present.

We appreciate the assistance of Allison Cohn, Tracy Saunders, and Martha Scherzer in editing the final version of this manuscript. For the design and layout of the document, we extend our thanks to Sue Koenigshofer and Erin Hancock of SCK Design.

♻ Printed on recycled paper

List of Acronyms

AIDS	Acquired Immune Deficiency Syndrome
CBR	Crude Birth Rate
CBD	Community–Based Distribution
CDC	Centers for Disease Control and Prevention
CDIE	Center for Development and Evaluation (USAID)
COPE	Client–Oriented Provider–Efficient
CPS	Contraceptive Prevalence Survey
CYP	Couple–Years of Protection
DHS	Demographic and Health Survey
EOC	Essential Obstetric Care
FIML	Full–Information Maximum Likelihood
FP	Family Planning
FPSS	Family Planning Support Services
GIS	Geographic Information Systems
HARI	Helping Individuals Achieve their Reproductive Intentions
HIV	Human Immunodeficiency Virus
IEC	Information–Education–Communication
IPPF	International Planned Parenthood Federation
KAP	Knowledge, Attitude, and Practice
MCH	Maternal and Child Health
MIS	Management Information Systems
MOH	Ministry of Health
NFPP	National Family Planning Program
NGO	Non–Governmental Organization
PEM	Protein–Energy Malnutrition
RH	Reproductive Health
SAM	Service Availability Module (of the DHS)
SCYP	Standard Couple–Years of Protection
SDP	Service Delivery Point
STD	Sexually Transmitted Disease(s)
TFR	Total Fertility Rate
WFS	World Fertility Survey

Table of Contents

Appendix

Chapter I

Overview of Evaluation

- Why Evaluate
- Objectives of this Manual
- Intended Users of the Manual
- Family Planning or Reproductive Health
- Scope and Focus of Evaluations
- Why Another Evaluation Manual
- Organization of this Manual

Chapter I

OVERVIEW OF EVALUATION

WHY EVALUATE

Evaluation is the application of social science research procedures to judge and improve the ways in which social policies and programs are conducted, from the earliest stages of defining and designing programs through their development and implementation (Rossi and Freeman, 1993). Evaluation results should inform program management, strategic planning, the design of new projects or initiatives, and resource allocation.

The evaluation of family planning programs includes both program monitoring and impact assessment. Monitoring is used to determine how well the program is carried out at different levels and at what cost; it tracks change that occurs over time in the resource inputs, production, and use of services. Impact assessment measures the extent to which this change can be attributed to the program intervention (cause and effect).

The results of program monitoring are indispensable for program management because they inform the manager whether the program is on track, where the problems are, and what unexpected results have occurred. Evaluation of the processes used in implementing the program allows for mid–course corrections. Indeed, the type of program monitoring described in Chapter III of this manual is integrally linked to management information systems (MIS), underscoring the close link between monitoring and management.

Evaluation results are also important inputs into strategic planning and program design. Measures of program performance, output, and population outcomes describe the current state of the demand for services and the program environment. Results linking inputs and activities to program outputs and changes at the population level serve to demonstrate what has worked in the past and to suggest potential directions for the future. Successful interventions can be scaled up or replicated in new program or project phases, whereas activities that do not produce results can be phased out. Moreover, evaluation can be used to explore why certain interventions did not work.

In short, those responsible for implementing programs and those who fund programs should require that evaluation be an integral part of any intervention. For maximum benefit, evaluation should be built into the program design from the start and provide data to managers over the life of the activity. Evaluation results will help administrators and managers to learn what they are doing right, identify shortcomings to be corrected, and make informed decisions about the future direction of their programs. In the current climate of budgetary constraints, evaluation results point to the most rational use of scarce resources—human and material—to achieve results.

OBJECTIVES OF THIS MANUAL

This manual prepares readers to:

- Differentiate between the main types of program evaluation, program monitoring and impact assessment;
- Critically evaluate the strengths and limitations of alternative methods for impact assessment;
- Assess and select the type(s) of evaluation most appropriate to a given setting (i.e., that answers the most important questions, yet is feasible/ practical);
- Identify appropriate indicators and sources of data for the evaluation; and
- Design an evaluation plan outlining study design(s), indicators, and sources of data that serves as a plan of action for subsequent implementation.

INTENDED USERS OF THE MANUAL

This manual is directed to health professionals with varied levels of training and experience in program evaluation, including:

- program administrators and managers,
- evaluation specialists, and
- donor agency personnel.

This manual assumes some training or experience in social science research techniques; it does not review basic procedures for data collection and analysis.[1]

FAMILY PLANNING OR REPRODUCTIVE HEALTH?

Until recently, the term "family planning" (FP) has been widely interpreted to mean "contraceptive services." Some programs also have provided a limited range of infertility services for couples unable to achieve pregnancy, but the bulk of service delivery has related to contraception.

However, in the period leading up to and after the 1994 Cairo International Conference on Population and Development, there was considerable pressure within the population community (especially from women's groups) to broaden the constellation of services to respond to other reproductive health (RH) concerns of women. These include safe pregnancy, sexually transmitted diseases and acquired immunodeficiency syndrome (STD/AIDS), breastfeeding, and women's nutrition, among others. In addition, adolescents and men have been identified as important target audiences for program interventions. The relative importance of the different services or target populations varies by country.

This manual draws on the extensive experience of the population community in evaluating family planning programs. Specifically, it provides guidelines relevant to developing an evaluation plan for a national family planning program, where the prime emphasis is on contraceptive services.

Are these principles equally applicable to programs in other areas of reproductive health? In terms of program monitoring, the logic and methodological techniques used are in many ways similar across types of FP/RH interventions. The differences lie in the selection of indicators, their measurement, and the definition of the target population, not the study designs or analytic techniques used.

Thus, the approach to evaluation described in this manual is generally applicable to other areas of reproductive health, either as vertical programs (e.g., a national program to promote breastfeeding) or as integrated programs with multiple components (e.g., breastfeeding, safe pregnancy, family planning, and prevention of STD/HIV). However, there are some notable differences, which are summarized in Chapter VII.

In terms of impact assessment, the logic and methodological techniques are also similar. In principle, the three "preferred methods" presented in Chapter IV are as applicable to other reproductive health interventions as to family planning. However, Chapter IV focuses almost exclusively on family planning for several reasons. First, because the question of impact in family planning programs has received more attention over the years than have other RH interventions, correspondingly greater attention has been paid to the development and refinement of evaluation methods for family planning than for other RH programs. The methodological advances described in Chapter IV have been developed in connection with efforts to assess the impact of family planning on fertility. No parallel work exists to date in relation to other RH outcomes. Second, one of the outcomes associated with family planning – fertility decline – can be measured validly from self–reported information on large–scale surveys. In contrast, there are major methodological difficulties in measuring other key outcomes in RH, such as maternal mortality, prevalence of HIV infection, and abortion rates. The emphasis on fertility in Chapter IV is not intended to diminish the importance of other reproductive health interventions. Rather, it reflects the fact that fertility lends itself more readily than most other reproductive health outcomes to accurate measurement based on self report from sample surveys.

[1] However, those interested in a useful background text are referred to Garcia–Nuñez (1992).

In short, this manual draws on the extensive evaluation experience from the field of family planning, much of which is applicable to the broader range of reproductive health interventions. It is intended as a companion volume to two other publications of The EVALUATION Project, the *Handbook of Indicators for Family Planning Program Evaluation* (Bertrand et al., 1994) and *Indicators of Reproductive Health Program Evaluation* (Bertrand and Tsui, 1995). These two documents provide a menu of indicators for evaluation interventions in the areas of family planning, safe pregnancy, breastfeeding, STD/HIV prevention, women's nutrition, and adolescent reproductive health services. By contrast, the current manual provides guidelines in designing an evaluation plan (that will incorporate those indicators) for monitoring and evaluating interventions.

SCOPE AND FOCUS OF EVALUATIONS

Evaluations vary greatly in scope and focus. For example, the target area may be defined as:

- the entire country;
- an entire region or state; or
- a specific city or location.

Evaluation can focus on different program components:

- inputs,
- processes,
- outputs, and
- outcomes.

Measurements can be taken at:

- the population level (e.g., among a random sample of the general population), or
- the program level (e.g., among clients or participants in a given program).

Different techniques are used to collect and analyze the data:

- quantitative, or
- qualitative.

The specific target population will vary in different settings and for different types of interventions:

- all women of reproductive age (e.g., family planning;
- all sexually active adults (e.g., integrated family planning — STD/HIV prevention); or
- youth aged 10 –19 (e.g., adolescent programs).

This manual focuses on the evaluation of programs that are national in scope, although many of the techniques can also be used to evaluate smaller–scale programs or projects. It addresses the key question of most program administrators and donor agency staff: has the program achieved its objectives in terms of change at the population level? It goes an additional step in asking to what extent the observed change is attributable to the program.

Although "results" are of tantamount importance, a comprehensive evaluation will also examine the processes involved in carrying out the program. Historically, family planning program evaluation has tended to focus heavily on quantitative outputs at the program level (e.g., number of new acceptors, couple–years of protection [CYP]) or outcomes at the population level (e.g., level of contraceptive prevalence, total fertility rate [TFR]). However, this approach to evaluation treats the program as a "black box." If the expected results are not obtained, it provides little insight into the reasons. One does not know, for example, what factors contributed to the poor results: inadequate access to service, poor quality of care at service delivery points (SDPs) in the system, lack of information among the target population, stockouts (lack of commodities) in the system? Similarly, if the program is successful, one has little knowledge of what contributed to the success.

In sum, a comprehensive evaluation will examine not only the quantitative outcomes that indicate progress toward program objectives; it will also evaluate the inner workings of the program in terms of functional areas (management, training, commodities and logistics, information, education and communication [IEC], research/evaluation) and service adequacy (access to services and quality of care). This manual focuses on both process and outcomes, with particular emphasis to the latter. Outcomes continue to be the primary concern of many program administrators and donor agencies, especially in an environment of shrinking resources and greater accountability.

WHY ANOTHER EVALUATION MANUAL

The population community has been interested in the evaluation of family planning programs since their inception in the 1960s. As a consequence, numerous methods of evaluation have been proposed and refined (Bogue, 1970; Reynolds, 1972; Sherris et al., 1985; Lloyd and Ross, 1989; U.N. Manuals, 1979, 1982, 1986; Garcia–Nuñez, 1992; Buckner et al., 1995). Why then do we need a new "how–to" document for evaluating FP programs? This document differs from previous evaluation texts in several ways:

- It addresses a need expressed with increasing frequency by donor agencies and program administrators: how to design an evaluation plan for a national program.
- It recognizes the shift in focus throughout the international donor community from family planning in the narrowly defined sense of contraceptive services to a broader definition that encompasses other aspects of reproductive health.

Figure I–1

Prototype Outline of an Evaluation Plan

What

Scope of the Evaluation:

- Goals and objectives of the program
- Conceptual framework that maps the linkages between inputs, processes, outputs, and outcomes
- Objectives of the evaluation:
  - ➢ program monitoring
  - ➢ impact assessment

How

Methodological Approach:

- Study design
- Indicators
- Data sources

Who, When, with What Funds

Implementation Plan:

- Individuals and institutions responsible for different parts of the evaluation
- Timetable for specific activities
- Budget

Why

Dissemination and Utilization of Results:

- Audiences
- Format and content

- It incorporates methodological developments and refinements produced to date by the USAID–funded EVALUATION Project into guidelines for field application.
- It updates the state–of–the–art in the most appropriate designs for assessing the impact of family planning programs (i.e., results attributable to the program) and it outlines the methodological issues in terms that are comprehensible to administrators and evaluation staff, yet satisfactory to the larger scientific community concerned with methodological rigor.
- It reflects the hands–on experience of EVALUATION Project staff in preparing evaluation strategies for specific countries; as such, it addresses the inevitable trade–off between scientific rigor and practical constraints.

ORGANIZATION OF THIS MANUAL

This manual is organized according to the elements of an evaluation plan, outlined in Figure I–1. An evaluation plan describes what will be done, how, by whom, when, with what funds, and why; it serves as a plan of action for implementation. Ideally, it will be prepared at the time the program is designed (or prior to its implementation). Where possible, the plan should include input from the program planners and managers who will work closely with the program and evaluation specialists who will carry out the evaluation research.

Each chapter addresses a section of the prototype evaluation plan, shown in Figure I–1. The information is intended to provide guidance in completing an evaluation plan for a FP program, tailored to a particular situation. The chapters cover the following:

Chapter II	Defining the scope of the evaluation
Chapter III	Methodological approach: program monitoring
Chapter IV	Methodological approach: impact assessment
Chapter V	Developing an implementation plan
Chapter VI	Disseminating and utilizing the results
Chapter VII	Adaptations to other reproductive health interventions
Chapter VIII	Summary: checklist of steps in designing an evaluation plan

Because the issue of demonstrating impact is key to evaluation methodology, we devote a separate chapter (IV) to this topic. In contrast to previous texts that provide a menu of eight "classical methods" for evaluating the impact of family planning programs on fertility, this section critically examines the alternatives and clearly advocates the use of three methods in particular. Readers will notice that the terminology in this section is more quantitative than in other chapters; nonetheless, we have attempted to present these concepts in a language comprehensible to those with a basic understanding of study design in social science research. Chapter IV can be read on two levels. For those with limited statistical background, the chapter identifies the three preferred methods ("preferred" because observed changes can be attributed to the program) and explains their strengths and limitations. It gives the reader a familiarity with the issues with which to make informed decisions regarding the type of evaluation to conduct, even if the reader is not directly responsible for the technical aspects of the work. For readers with more statistical background, Chapter IV provides the rationale for labeling three specific methods as "preferred" and summarizes the statistical estimation approaches and procedures entailed by these methods.

Program administrators and donor agency staff may surmise at first glance that Chapter IV is intended for evaluation specialists only. However, it is precisely these two categories of health professionals who are often under pressure to "demonstrate impact." Readers who might otherwise be discouraged by the quantitative language in Chapter IV are encouraged to focus on the concepts, not the statistics.

Chapter II

Defining the Scope of Evaluation

- Determining the Program Goals and Objectives
- Describing How the Program "Should" Work
- Establishing the Objectives of the Evaluation
- Outlining the Scope of the Evaluation

Chapter II

Defining the Scope of Evaluation

DETERMINING PROGRAM GOALS AND OBJECTIVES

Defining the Program

Different people have different ideas of what constitutes a "program." In this manual we are primarily interested in the evaluation of programs at the national level. However, many of the techniques described herein are equally applicable to smaller–scale interventions. The box below defines the terms "program" and "project" as they are used in this manual.

In most developing countries today, there are multiple organizations delivering family planning and related reproductive health services. Whereas each organization should monitor its own performance and results, high–level decision makers and donor agency staff are equally interested in evaluating changes at the national level (e.g., in contraceptive prevalence) that result from the collective efforts of all programs and sources providing relevant services. The conglomeration of the relevant services is often referred to as "the national program," even in the absence of an official coordinating entity. Thus, in a given country the "national program" might include the Ministry of Health (MOH) services, the International Planned Parenthood Federation (IPPF) affiliate, other non–governmental organizations, a subsidized social marketing program, and the for–profit commercial sector.

Defining Goals and Objectives of the Program

In the ideal case, the program will have well–articulated goals and objectives.

The goal is a statement, usually general and abstract, of a desired state toward which a program is directed (Rossi and Freeman, 1993). A goal may be stated in terms of the activities of an entire agency, or it can be specific to a particular country. For example, one of the goals of USAID is:

- stabilizing world population and protecting human health (USAID, 1995).

Examples of country–specific goals include:

- sustained improvement in the standard of living, and
- political, social, and economic empowerment.

Objectives, by contrast, are specific, operationalized statements detailing the desired accomplishments of a program (Rossi and Freeman, 1993). Examples of population–level objectives relevant to family planning include:

- to reduce the total fertility rate to 4.0 births by Year X;
- to increase contraceptive prevalence to 50% by Year X; and
- to achieve a median interval of 36 months between births by the end of the five–year project.

Some countries have quantifiable objectives for their programs (as in the example directly above). In other cases the stated objective will indicate the desired direction of change without quantifying the magnitude of change (e.g., to increase contraceptive prevalence over the life of the program). Either case can be "put to the test," although the results are less ambiguous when the expected level of increase has been quantified.

National programs generally have goals that specify population–level results such as changes in contraceptive prevalence, infant mortality, total fertility, and so forth. Institutional programs (e.g., the program of the IPPF affiliate) and projects within programs often share these same goals, but realize that they will not be able to measure their contribution to the expected change in population–level outcomes; thus, their objectives are often stated in terms of expected results of activities at the program level.

Evaluators are sometimes faced with a situation where the objectives of the program are not stated in measurable terms. In this case, part of the process of establishing the objectives is to assist in stating the program objectives in terms that lend themselves to evaluation; that is, to "operationalize" the objectives.

DEFINITION: Program

The different types of organized activity common to family planning institutions can be classified as follows. This manual focuses primarily on the first definition, but many of the evaluation principles and tools apply across all three.

National Family Planning Program

Definition: All organized activities designed to promote family planning in the public, private voluntary, and commercial sectors in a given country.

This array of activities may be coordinated by a government or para–statal body such as the Zimbabwe National Family Planning Council or (BKKBN) Badan Koordinasi Keluarga BerencanaNasional,the National Family Planning Coordinating Board in Indonesia. More commonly, different organizations such as the Ministry of Health (MOH), the International Planned Parenthood Federation (IPPF) affiliate, the commercial private sector, and other non–governmental organizations (NGOs) work in a somewhat autonomous manner but share one or more common objectives: generally, to improve social conditions and human welfare by assisting couples to exercise their right to choose the number and timing of their children, enhance the health of women and children, and/or slow population growth. Ideally, there will be a formal document to define the national goals and operational mechanisms for achieving them, as a basis against which to evaluate the program.

The Program of an Institution (Institutional Program)

Definition: The collection of activities that a given public or private sector organization implements in pursuit of its FP objectives.

Examples include the MOH program, the IPPF affiliate's program, and the social marketing program (where it is not subsumed by another institution). These programs often have more specific objectives than the national program. For example, in contrast to the national program that is aimed at increasing contraceptive prevalence overall, the public sector may be specifically concerned with providing family planning methods to low–income populations.

DESCRIBING HOW THE PROGRAM "SHOULD" WORK

Program Components

In its broadest conceptualization, a family planning program can be viewed in terms of four distinct elements: inputs, process (or activities), outputs, and outcomes (Veney and Gorbach, 1993).

- Program inputs refer to the set of resources (i.e., personnel, facilities, space, equipment and supplies, etc.) that are the raw materials of the program.
- Program processes refer to the set of activities in which program inputs are utilized in pursuit of the results expected from the program. Program processes include all of the service delivery operations[2] (management, training, commodities and logistics, information–education–communication, and research and evaluation) that the program conducts in order to provide family planning services.
- Program outputs are the results obtained at the program level through the execution of activities using program resources. There are three types of program outputs:
  - ➤ functional area outputs, such as the number of persons trained and the number of IEC talks;
  - ➤ service outputs, such as access to services and quality of care; and
  - ➤ service utilization, such as couple–years of protection (CYP), and the number of new acceptors.
- Program outcomes are the set of results expected to occur at the population level due to program activities and the generation of program outputs. These may be divided into two

[2] Also known as "functional areas."

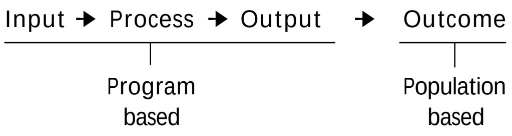
components: intermediate outcomes,[3] and long–term outcomes.

- Intermediate outcomes are the set of results at the population level that are closely and clearly linked to program activities and program level results. The most familiar intermediate outcome of a family planning program is contraceptive use. Changes in intermediate outcomes generally occur within 2–5 years of program inception.
- Long–range outcomes refer to the set of results at the population level that are long–term in nature and are produced through the action of intermediate outcomes. Examples of long–term outcomes of family planning programs are changes in fertility and maternal–child health status. Although health and fertility rates can change abruptly in response to external forces, there is generally a considerable time lag (5–10 years) between the inception of the program and the observance of change in these rates.

Inputs (program resources) are fed into processes (program activities), which in turn produce outputs (program results) and ultimately outcomes (changes in population behavior).

Figure II–1

Relationship between Program Components



The first three —inputs, processes, and outputs— relate to activities and results at the program level. Inputs, processes, and outputs are measured with program–based or facility–based data (see Figure II–1). Program–based data come from routine data collection (e.g., service statistics, client and other clinic records, administrative records, commodities shipments, sales) as well as information that is collected on–site where services are delivered (e.g., provider surveys, observation of provider–client interaction, retail audits, mystery clients) or from a follow–up study of clients.

DEFINITION: Project

Project (within an institution)

Definition: A specific set or cluster of activities with specific objectives that contribute to the overall objectives of the institution. The different projects of an institution, which may be funded from different sources, collectively constitute its program. A project often has specific attributes in terms of:

- target population (e.g., adolescents, men);
- method promoted (e.g., long–acting methods);
- type of service delivery mechanism used (e.g., community based distribution [CBD], traditional midwives); and
- constellation of other health and social services provided concurrently (e.g., prenatal).

One may also hear reference to the “population program” of a donor agency. For example, multi–lateral agencies such as the United Nations Population Fund (UNFPA) and IPPF, and bi–lateral development agencies of certain countries (USA, Japan, Germany, England, Sweden, Canada, among others) support a range of activities in the population sector in different countries. A donor agency often provides funding for one or more projects within an institution, and in isolated cases for the entire program of the institution. Although the activities funded by a given donor may differ markedly in nature (e.g., conducting a census, supporting observational travel for key decision–makers, purchasing commodities), the portfolio of population–related activities of a donor agency is often referred to as its “population program.” This manual addresses the evaluation of donor funding to the extent that these funds are channeled into a family planning program or project in a specific developing country.

Programs and projects are closely related. The set of multiple projects of a given institution collectively constitute the institutional program. The various institutional programs collectively make up what we term “the national program.” In turn, these programs are generally funded in part by the government and possibly one or more donor agencies.

[3] Alternatively, these are referred to as “short– to mid–range outcomes.”

Outcomes (intermediate and long–term changes at the population level) are measured with population–based data collected from the catchment area or social group that the program seeks to benefit. This may be a country, a region, or a particular sub–group of the population (e.g., adolescents).

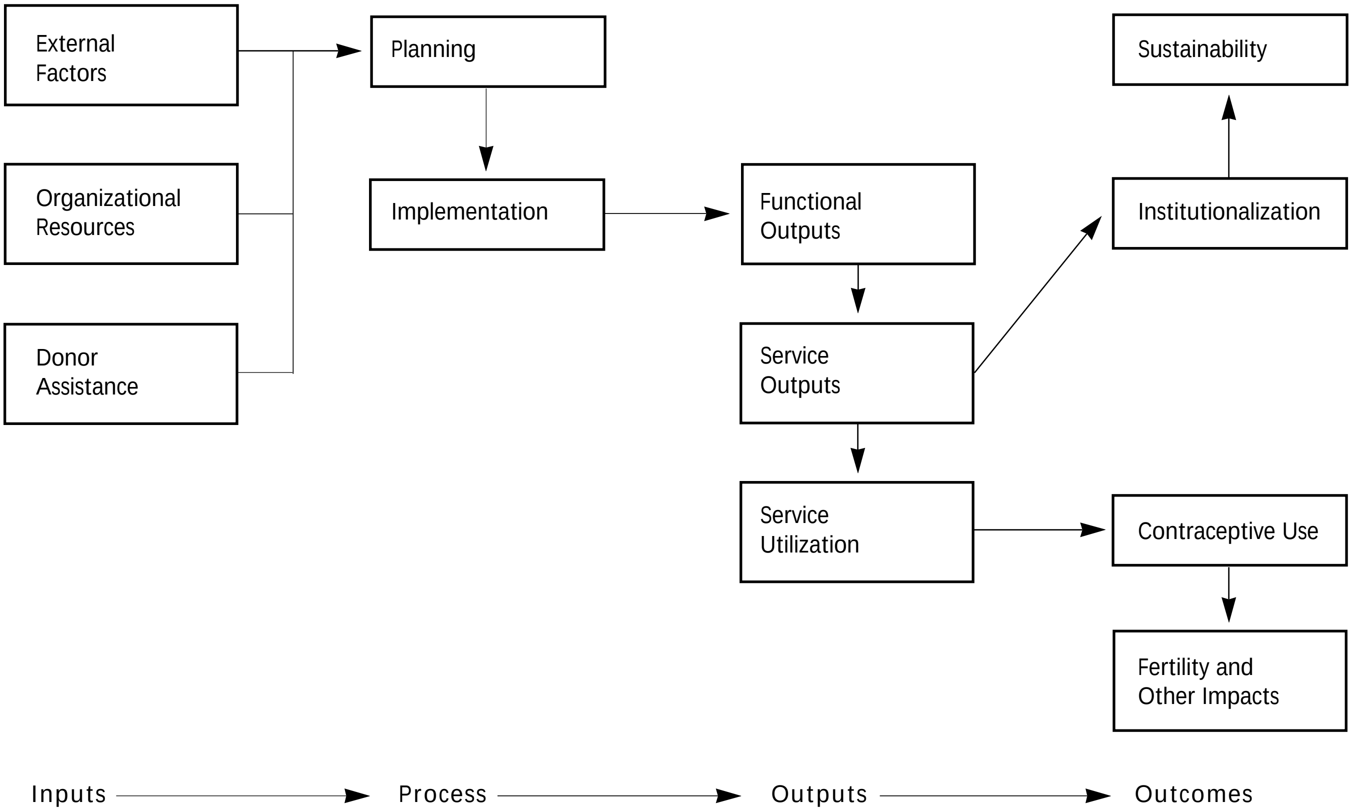
Describing Causal Linkages

Once the program objectives are established, it is important to define how the activities in the program are expected to achieve these objectives. The expected causal sequence is shown in its simplest form in Figure II–1 on the previous page. The flow diagram (conceptual model) in Figure II–2 indicates the content (factors, activities, results, etc.) that are included under each of the broad categories of input, process, output, and outcome.

Figure II–2 is useful in that it shows the relation between program components and the terms input–process–output–outcome. The diagram in Figure II–2 is appealing because of its simplicity; however, it masks the complexity of real programs. For example, "planning" and "implementation" entail an array of interrelated activities from the different functional areas (management, training, commodities/logistics,

Figure II–2

Components of Program Impact



Source: Tsui, A.O. and P. D. Gorbach, 1996. *Framing Family Planning Program Evaluation: Cause, Logic and Action.* The EVALUATION Project, University of North Carolina at Chapel Hill.

and IEC). In mapping out the linkages between the program and the expected results, it is important to identify the (often multiple) components of the "processes."

Figure II–2 shows only part of the picture: the contribution of program activities (the "supply side") to the desired results. In the real world, the success of programs is also determined by demand factors (social, economic, and other non–program variables that affect demand for the service in question). Figure II–3 illustrates a full conceptual framework used by The EVALUATION Project to describe the pathways by which demand factors and supply factors both contribute to program results and population–level outcomes. In Figure II–3 the demand factors are shown in the upper left side; the supply factors are found on the bottom left side and further elaborated in Figure II–4. The text in the box below on the "Conceptual Framework" further explains these relationships.

Figures II–2 (showing the program components in terms ofinput–process–output–outcome) and Figure II–4 (detailing the family planning supply environment) are closely related, and in both cases they exclude household demand factors (i.e., whether people want family planning services).

These diagrams are useful as road maps to visualize how one expects to get from point A to point B. Furthermore, they allow practitioners and evaluators to clarify how the multiple, diverse ingredients (such as program leadership, policy regulations, management style, quality of staff training, range of products or contraceptives available, frequency of stockouts, quantity and quality of IEC, and quality of care) come together in a given program, and to conceptualize them in a systematic way that facilitates evaluation.

The model in Figures II–3 and II–4 describes the expected linkages for a vertical family planning program aimed at the reduction of fertility in the long run. However, as described in Chapter I, the "model" for family planning programs has evolved considerably in recent years to include other reproductive health services (related to safe motherhood, breastfeeding, sexually transmitted disease [STD], post–abortion services). The

CONCEPTUAL FRAMEWORK: The Impact of Family Planning Programs on Fertility

The conceptual framework is the basis for identifying appropriate program indicators and specifying the pathways by which program inputs produce program outputs and ultimately changes in the behavior of the target population. A conceptual framework describing the linkages between family planning program inputs and fertility change is shown in Figures II–3 and II–4. The Handbook of Indicators for Family Planning Program Evaluation outlines key indicators for each box on the diagram. The framework can (and should) be modified to evaluate family planning programs with different objectives or programs that strive to achieve these objectives through a different constellation of services.

The framework recognizes that fertility and other impacts are the consequence of both the demand for and supply of family planning services. Demand for children and demand for family planning services are affected by a number of political, socioeconomic, cultural, and individual factors. Thus, an increase in the availability of family planning services is more likely to translate into higher levels of use in a country where these other factors exert a positive (rather than negative) influence on demand.

The family planning supply environment (Figure II–4) is also shaped by the political and administrative systems within which the program operates. Political support for the family planning program, funding of the program, and the legal and regulatory environment affect program organization and success. Inputs to the family planning program in the form of personnel, facilities and space, equipment and supplies, etc., are transformed through program activities. These program activities consist of the planning and implementation of the principal family planning program functions: management, training, distribution of contraceptives and related supplies, IEC efforts, and research and evaluation. Collectively, the results in these functional areas create the principal program outputs—accessibility, quality, and well–regarded family planning services. These outputs attract clients to the program and, jointly with demand for family planning, determine the impact of the program on the target population.

The total costs of inputs used to produce particular program outputs and outcomes may be calculated. For example, the unit costs of personnel, supplies, and other inputs may be multiplied by amounts used, summed, and then related to outputs or the services produced by the program.

Figure II–3

Conceptual Framework of Family Planning Demand and Program Impact on Fertility

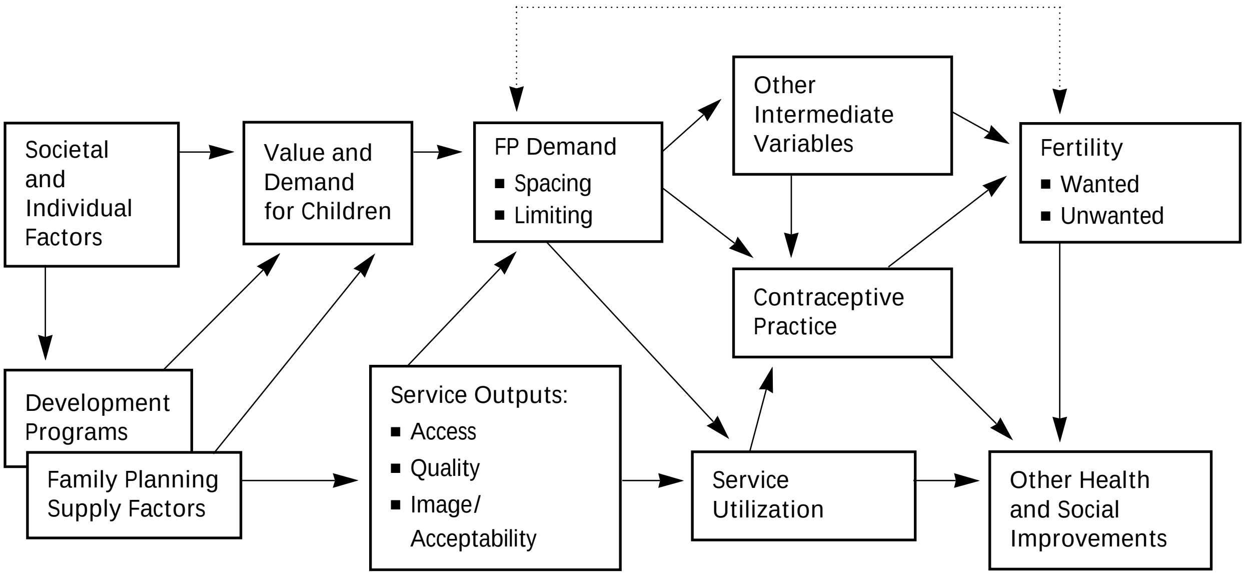
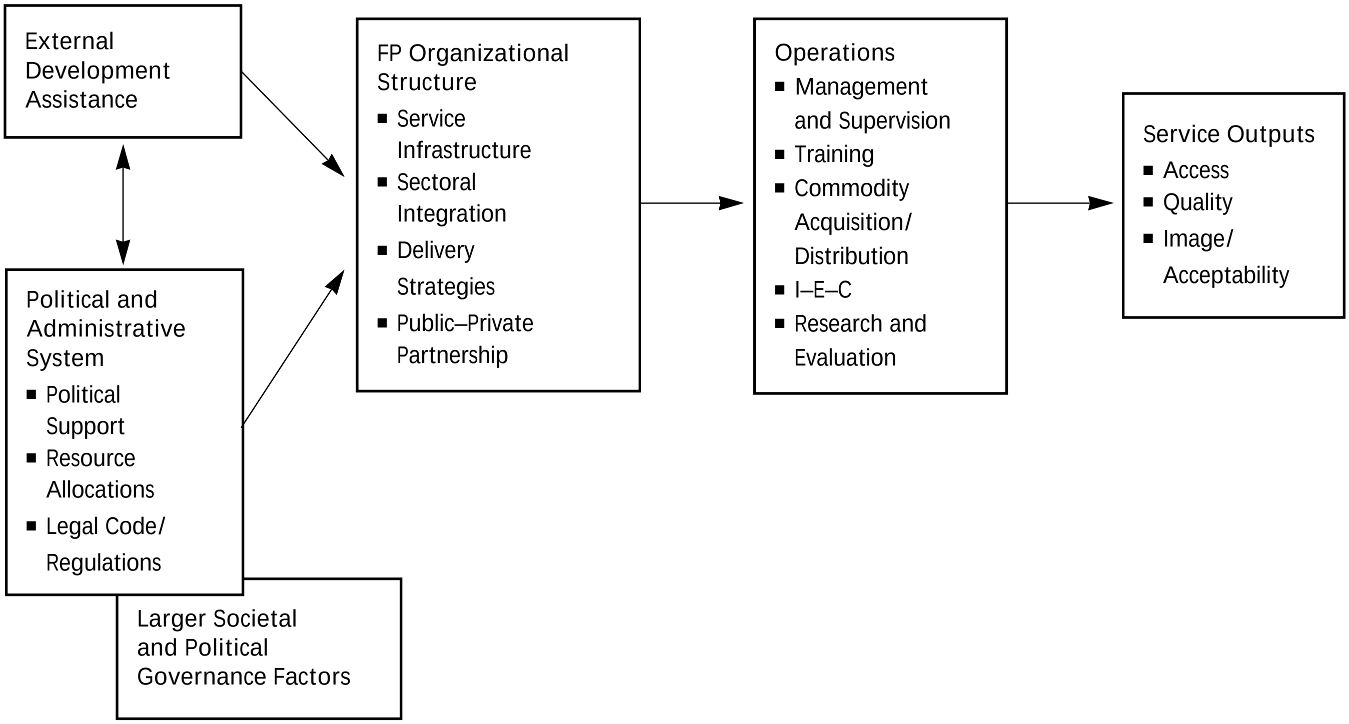


Figure II–4

Conceptual Framework of Family Planning Supply Factors


External Development Assistance
Political and Administrative System
■ Political Support
■ Resource Allocations
■ Legal Code/ Regulations
Larger Societal and Political Governance Factors
FP Organizational Structure
■ Service Infrastructure
■ Sectoral Integration
■ Delivery Strategies
■ Public–Private Partnership
Operations
■ Management and Supervision
■ Training
■ Commodity Acquisition/ Distribution
■ I–E–C
■ Research and Evaluation
Service Outputs
■ Access
■ Quality
■ Image/ Acceptability

expected outcomes relate not only to fertility but also to maternal and child health status and the satisfaction of individual reproductive intentions (as measured by the HARI index[4]).

ESTABLISHING THE OBJECTIVES OF THE EVALUATION

Alternative Types of Evaluation

There are different types of evaluation, each with a different purpose, as outlined in Figure II–5. In designing an evaluation strategy, the evaluator needs to identify the key question(s) that he/she wishes to answer and thus the type of evaluation to conduct.

One of the main objectives of this manual is to clearly differentiate program monitoring and impact assessment. As reflected by the illustrative questions in Figure II–5, program monitoring addresses a number of different questions, one of which is: did change occur? However, without impact assessment, one can not answer the question: did change occur because of the program?

Figure II–5

Types of Evaluation

Type	Question(s) Addressed
Needs Assessment	What should the program include and how can it best be delivered to meet the needs of the target group?
Program Monitoring	Inputs ■ Were inputs (e.g., equipment, commodities, materials, personnel) made available to the program in the quantities and at the times specified by the program plan? Processes ■ Were the scheduled activities carried out as planned? ■ How well were they carried out? Outputs ■ Did the expected changes occur at the program level, in terms of: ➢ access to services ➢ quality of care ➢ service utilization Outcomes ■ Did the expected change occur at the population level (not necessarily attributable to the program)? Costs ■ What was the incremental cost of: expanding an activity, producing a higher unit of output, and achieving the change that occurred?
Impact Assessment	What and how much change occurred (at the program– or population–level) that is attributable to the program?

[4] HARI is an acronym for "Helping Individuals Achieve Their Reproductive Intentions" (Jain and Bruce, 1994). It measures the extent to which members of the target population achieve their reproductive intentions (e.g., to have another child, to avoid further pregnancy).

The remainder of this chapter compares and contrasts program monitoring versus impact assessment.

Purposes of Program Monitoring

Monitoring[5] refers to a varied set of evaluation techniques, all of which measure some aspect of program performance. There are two main purposes of program monitoring:

- to improve programs by identifying those aspects that are working according to plan and those that are in need of mid–course corrections, and
- to track changes in the services provided (service outputs) and the desired results.

Monitoring includes measuring the current status and change over time in any of the program components.

At the program level:

- Inputs
- Outputs: Functional outputs, Service outputs (or service adequacy), Service utilization

[5] The monitoring of activities related to program–level variables is also called "process evaluation."

Figure II–6

Using the Conceptual Framework to Identify Issues to Address in Program Monitoring

Program Component	Quantity of Component	Quality of Component	Cost of Component
Inputs	What types and level of resources were allocated to this intervention?	Were qualified personnel available to implement activities?	What was the unit cost of each resource? the total cost of each resource? the total cost of the program?
Outputs Functional areas (e.g., training)	How many persons were trained, by category of personnel?	Were trained staff able to perform tasks competently 6 months post–training?	What was the cost per participant–day of training?
Outputs Service outputs	Access: Did the number of SDP's providing services increase?	Quality: Did the quality of care improve over time?	What was the added cost of increasing the numbers of SDPs? of improving the quality of care?
Outputs Service utilization	Did the number of new acceptors or CYP increase over time?	Has percent of clients returning for follow–up appointments increased?	What was the added cost associated with the increase in new acceptors? with the increase in CYP?
Outcome Intermediate outcomes	Was there change in the key behavior (e.g., contraceptive prevalence) among the target population?	Was there a change in the key behavior (e.g., receiving quality of care) in the target population?	What was the increase in costs associated with the change in contraceptive prevalence?
Outcome Long–term outcomes	Did women achieve their reproductive intentions?	Did fertility rates change over time?	What was the cost of achieving the fertility change?

At the population level:

- Outcomes: Intermediate outcomes (e.g., contraceptive prevalence), long–term outcomes (e.g., total fertility rate).

Some readers may be surprised to see change at the population level categorized as monitoring rather than impact assessment, since in many programs the "impact sought" is a decline in fertility, mortality, or morbidity. However, it is not possible to attribute cause and effect or determine the percent of change attributable to the program based on simple monitoring of trend data. Even when the variables in question refer to the ultimate changes that the program desires to bring about, the tracking of these trends falls under the broad category of program monitoring (in the absence of a study design or statistical analysis that establishes causality).

Ideally, program monitoring will include both quantitative and qualitative research techniques. Data collection may include any of the standard techniques used in social science research (surveys, focus groups, in–depth interviews, observation, key informants, etc.) in addition to the analysis of program data (e.g., service statistics).

Figure II–6 illustrates how the conceptual framework can be used to identify and organize the key questions to be addressed regarding quantity, quality, and cost of programs. For simplicity, only one item per box is shown in Figure II–6. In an actual evaluation, one would select the boxes of particular interest and identify indicators (variables) to measure how well the program was doing in each respect. A comprehensive program evaluation would attempt to cover a number of these boxes. By contrast, a special study might focus on a single one, such as quality of care at a set of SDPs.

Program monitoring is by far more common than impact assessment in evaluating national family planning programs.[6] At the program level, administrators use the trends in program data (number of new acceptors/clients, volume of CYP, number of clinic visits by purpose of visit, number of contacts with adolescents, and so forth) as a means of assessing progress and identifying areas in need of improvement. Moreover, they track population–level trends over time (in contraceptive prevalence, method mix, total fertility rate, median length of birth intervals, etc.) to assess progress toward intermediate and long–term objectives. In general, although monitoring of outputs and outcomes does not in itself demonstrate cause and effect, there is often an assumption of plausible association.

Purpose of Impact Assessment

The purpose of impact assessment is to measure the degree of change attributable to a given program or intervention. In contrast to program monitoring, which simply tracks change, impact assessment addresses the question of causality.

The word "impact" is frequently heard in organizations whose mandate is to bring about behavioral change. Often program personnel will use this term to describe the desired change they hope to achieve through a given intervention, and as such they are describing their own conceptual framework that links inputs to processes to outputs and outcomes. For example, the social marketing project staff may speculate on the impact of their promotional activities on sales. The IEC staff hopes its current national campaign will have the desired impact of bringing clients into contact with the services.

It is easy to describe this type of cause and effect relationship, but it is generally more difficult to demonstrate impact empirically. The reason, as mentioned above, is that even when the desired change occurs, it cannot necessarily or exclusively be attributed to the program intervention. For example, the director of the national family planning program may cite the increase in contraceptive prevalence over time as "proof of the impact of the national program," and in lay terms many will accept this statement at face value. In technical terms, however, the word "impact" denotes that the evaluation is based on a study design demonstrating not only change, but cause and effect.

[6] There are numerous operations research studies that use quasi–experimental designs to link changes in some aspect of service delivery with a change at the program or population level. Many of these studies are designed to demonstrate cause and effect. However, they are generally performed in a specific geographical area, and thus they do not constitute an evaluation of the national program.

This distinction is illustrated in Figure II–8. The diagram is included to emphasize the point that an evaluation may demonstrate the expected change, but this is not necessarily attributable to the program.

In sum, family planning evaluation often consists of program monitoring only. However, when one assesses impact, the analysis should include measures of program performance. Indeed, the variables used to assess impact (and thus the information collected) are the same as the indicators used for program monitoring. What differentiates the two are the evaluation design and analytic techniques used.

This manual encourages evaluators to distinguish between program monitoring and impact assessment in designing evaluation plans and reporting results. Moreover, it calls for renewed effort by members of the evaluation community to conduct impact assessment in the rigorous sense of the term, and it provides methodological guidance for doing so.

As a preview to the more detailed arguments in Chapter IV, Figure II–9 summarizes the "preferred methods" for impact assessment and the data needed for each.

Factors in Deciding When to Assess Impact

In practical terms, the question is NOT whether to do program monitoring OR impact assessment. Rather, it is when to carry out impact assessment in addition to program monitoring. The decision will be based in large part on the following issues:

- the current stage of implementation of the program in question;
- the availability of resources for additional data collection and analysis; and
- the need to demonstrate impact (e.g., to justify continued funding).

Current Stage of Implementation of the Program

The evaluation plan for a given program should be developed from the start. (In the case of USAID, this is usually the beginning of a five–year funding cycle.) Once a program is underway, it is generally too late to apply retroactively an evaluation design

RESULTS FRAMEWORK: USAID's Approach to Program Monitoring

USAID currently monitors program performance using a results framework, as shown in the example in Figure II–7 from the USAID mission in Morocco.

Under this system, a given mission develops one or more strategic objectives. The strategic objective is the most ambitious result for which the operational unit, along with its partners, is willing to be held responsible. It forms the standard on which its performance will be judged. In the population and health sectors, strategic objectives range from decreased fertility and mortality to increased use of selected family planning, child survival, and/or reproductive or maternal health services.

A results framework is then developed for each strategic objective. This framework illustrates the causal pathways that lead to the achievement of the objective(s), as well as the results needed at preceding levels for their achievement. This framework is also useful in communicating the underlying premises of the strategy. The results framework forms an essential part of the strategic plan that must be developed by each overseas mission or operating unit that uses program funds. In addition to the strategic objective(s) and results framework, the plan outlines the approaches to be used in achieving the objectives, the indicators for measuring results, and the frequency of reporting for each.

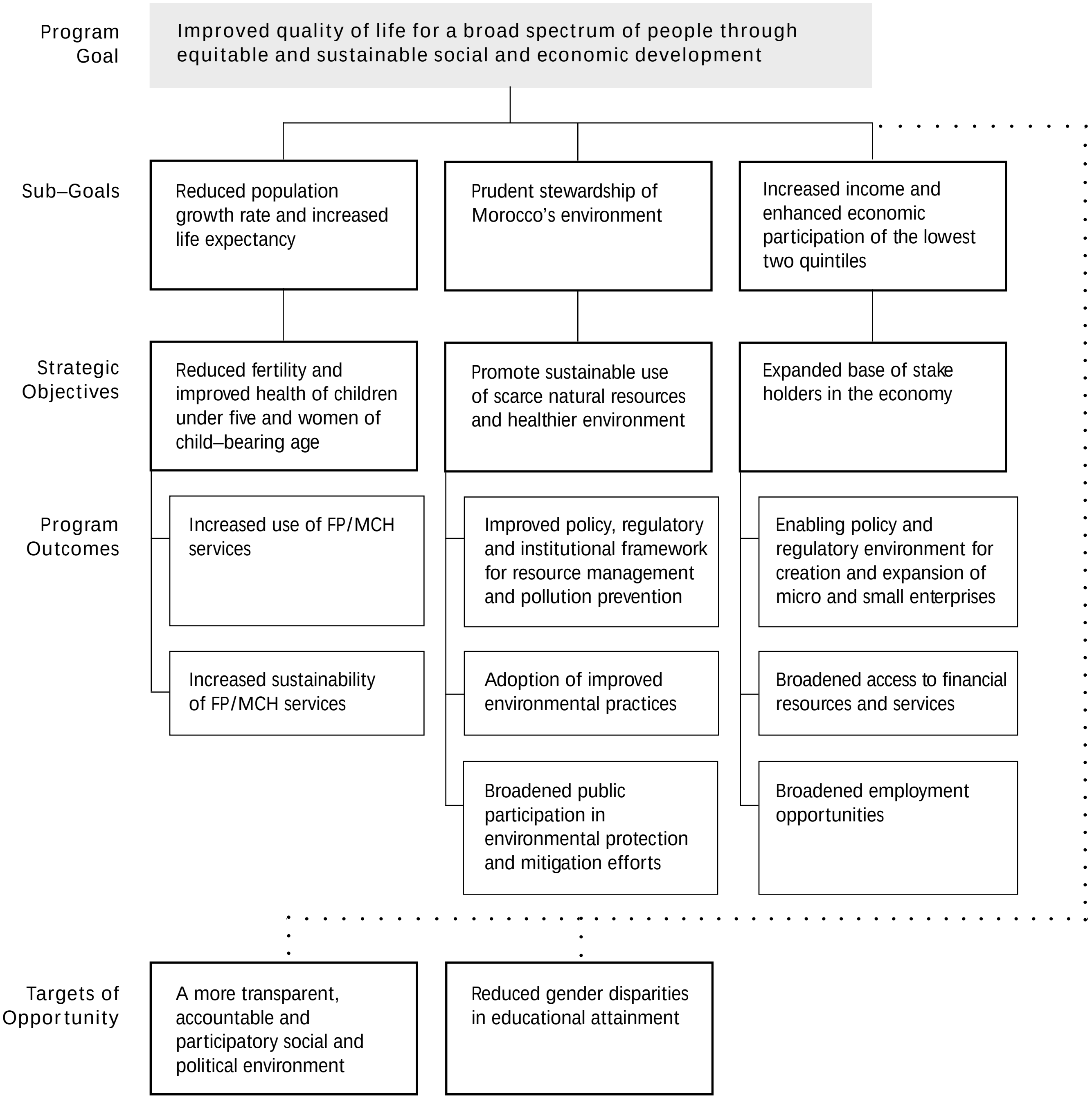
Each strategic objective and result must have at least one performance indicator. These indicators must be clear, precise, and objectively measurable. These indicators are measured at a baseline and subsequently at one or more points during the project cycle. In addition, the results framework should include any key results produced by other development partners (e.g., such as NGOs, the host country government, other donors and customers).

Operating units document their progress toward achieving the stated results in an annual report that contains both a narrative portion and performance indicator tables.

Figure II–7

Example of a Monitoring System for a National Program

USAID/Morocco Program


Program Goal
Improved quality of life for a broad spectrum of people through equitable and sustainable social and economic development
Sub–Goals
Reduced population growth rate and increased life expectancy
Prudent stewardship of Morocco's environment
Increased income and enhanced economic participation of the lowest two quintiles
Strategic Objectives
Reduced fertility and improved health of children under five and women of child–bearing age
Promote sustainable use of scarce natural resources and healthier environment
Expanded base of stake holders in the economy
Program Outcomes
Increased use of FP/MCH services
Increased sustainability of FP/MCH services
Improved policy, regulatory and institutional framework for resource management and pollution prevention
Adoption of improved environmental practices
Broadened public participation in environmental protection and mitigation efforts
Enabling policy and regulatory environment for creation and expansion of micro and small enterprises
Broadened access to financial resources and services
Broadened employment opportunities
Targets of Opportunity
A more transparent, accountable and participatory social and political environment
Reduced gender disparities in educational attainment

Source: USAID/Morocco Country Program Strategy 1995–2000 (April 1995) that allows a credible measurement of impact. Moreover, developing an evaluation plan at the start of the program cycle ensures that the necessary data are being or will be collected.

For evaluators asked to design an evaluation plan before the program begins, the "phase of implementation" is not a problem. Those given the task when the program (or current funding cycle) is already underway are more constrained in their options. Under less than optimal conditions, the evaluator has the following options:

- limit the evaluation to program monitoring only;
- attempt to demonstrate linkages between program components and output/outcome measures, even if it is not possible to show cause and effect unequivocally (see Chapter IV); or
- identify stronger designs that can be used in subsequent phases of activity.

Availability of Existing and Potential Data Sources

The availability of data will also enter into the decision to conduct impact analysis. The data requirements for the three "preferred approaches" differ somewhat, as briefly summarized in Figure II–9 and described in detail in Chapter IV. This is often the determining factor in the decision.

OUTLINING THE SCOPE OF THE EVALUATION

Finally, it is important to present the decisions of what to evaluate in written form. As mentioned in Chapter I, this description of the scope of the evaluation should include:

- Goals and objectives of the program
- Conceptual framework that maps the linkages between inputs, processes, outputs, and outcomes
- Objectives of the evaluation
  - ➤ program monitoring
  - ➤ impact assessment

There is no need for lengthy prose in presenting these three topics in an evaluation plan (although it is important to make the distinction between program monitoring and impact assessment). Rather, what counts is conceptual clarity.

With a clear description of the type(s) of evaluation to conduct, one proceeds to the specifics of methodological approach. The next chapter (III) provides guidelines for program monitoring. Certain concepts in Chapter III (e.g., selection of indicators, data sources) are equally applicable to impact assessment. However, the issues of study design for measuring impact are sufficiently different that they are discussed separately in Chapter IV.

Figure II–8

The Continuum of Types of Evaluation

Process	Results	Impact
How well did the program work: ■ What and how many activities were implemented? (quantitative assessment) ■ How well were they implemented? (qualitative assessment)	Did the expected change occur: ■ at the program level? (outputs) ■ at the population level? (outcomes)	To what extent can the change be attributed to the interventon? Based on a theoretical model and demonstrated by: ■ an experimental or quasi–experimental design; or ■ multilevel longitudinal analysis.
<—— Monitoring Program Performance ——>		<—— Assessing Impact ——>

Figure II–9

Preferred Methods for Assessing Impact in Family Planning Programs

Previous texts and manuals on evaluating the impact of family planning programs have described a number of methods or approaches (United Nations, 1979, 1982, 1985; Hermalin, 1982; Sherris et al., 1985; Buckner et al., 1995). However, the consensus developed by The EVALUATION Project is that only three approaches adequately demonstrate causality (i.e., that observed change is attributable to the program). Described in fuller detail in Chapter IV, these three preferred approaches are as follows:

Randomized experiment: the "pretest/posttest control group design"

This design is widely viewed as "the gold standard" for evaluating impact, because (when implemented appropriately) it answers the question: "what would have happened in the absence of the program?" By comparing the change that occurs in the experimental versus control populations, one can measure the amount of change atributable to the program ("net" of confounding factors). The major limitation to this method relates to the feasibility and political acceptability of conducting experiments.

Data needs: One must obtain data from two groups: those who do and do not receive the intervention; moreover, subjects (or groups of subjects, such as villages) must be randomly assigned to the experimental versus control group. Data (either service statistics or survey data) are collected both "pre" and "post" intervention for the two groups.

Although this design is often criticized for being difficult to implement, one promising approach is the "siting" of interventions. Specifically, if one is at the beginning of a project cycle that will entail new interventions or a pilot study prior to full–scale expansion, it may be possible to allocate randomly the facilities or areas that do and do not receive the intervention, and subsequently compare the results from the two groups. This approach may be particularly useful where resources are too limited to begin the intervention in all areas simultaneously.

A quasi–experimental design: "non–equivalent control group design"

This approach is similar to the pretest–posttest control group design described above, except that the subjects (or administrative units such as villages) are NOT randomly assigned to experimental versus control groups. Rather, the researcher selects or assigns villages based on their similarity of socio–demographic or other key characteristics. Otherwise, the data needs are the same as above.

Multilevel longitudinal regression models

This methodology uses statistical techniques common in social sciences, applied to the types of data available for family planning program evaluation. In brief, it is designed to demonstrate empirically that program inputs resulted in changes at the program level (e.g., more and better services) AND that these outputs produced a change in the desired behavior, such as contraceptive practice (i.e., the intermediate outcome). In some cases, the long–term outcome is used instead or as well (e.g., fertility rate).

Data needs: The data requirements for this type of analysis include surveys of the target population (which can be the country as a whole) at two points in time conducted in the same sample clusters, as well as data on the service delivery network at the same two points. (In some cases, it is possible to do similar analyses from a study at one point in time if information on key variables can be reconstructed retrospectively for a period several years earlier.) DHS–type surveys provide information on the target population. The data on the service delivery network can be obtained from the service availability module (SAM) of the DHS and/or a Situation Analysis study. Few countries currently have household and facility data for the same set of sample clusters. However, many countries already have one DHS survey with a service availability module and are planning for a second DHS; in this case, the addition of the second SAM would provide the necessary data for this approach.

Note: The above methods are listed in an order intended to facilitate presentation, not necessarily in order of preference. The conditions under which the different methods would be preferred are discussed in Chapter IV.

Chapter III

Methodological Approach: Program Monitoring

- Clarifying the Primary Purpose of Monitoring
- Identifying the Components to be Monitored
- Defining Relevant Indicators
- Identifying Sources of Data
- Designing a Format for the Presentation of Results
- Summarizing the Methodological Approach

METHODOLOGICAL APPROACH: PROGRAM MONITORING

This chapter assumes that the evaluator has completed the first step in developing an evaluation plan: defining the scope of the evaluation (i.e., to monitor program performance only, to measure impact only, or to do both).

If the decision includes program monitoring, the next steps in developing an evaluation plan consist of defining the :

- primary purpose of monitoring,
- components (aspects) of program to monitor,
- study design(s),
- indicators,
- sources of data, and
- format for presenting results.

CLARIFYING THE PRIMARY PURPOSE OF MONITORING

As mentioned in Chapter II, program monitoring has two main purposes:

- to improve programs by identifying those aspects that are working according to plan and those that are in need of mid–course corrections, and
- to track (and demonstrate) results at the program or population level.

For example, certain evaluation techniques are designed specifically to improve performance. In an effort to enhance quality of care in family planning programs, a number of countries have experimented with the Client–Oriented Provider–Efficient (COPE) technique, which is a self–assessment tool designed for use at the local level (AVSC International, 1995). The data are NOT aggregated to a higher level, but rather are analyzed by the service providers to identify changes that can take place at the local level to address problems identified by the exercise. This qualitative technique represents a promising approach to improving programs from the bottom upward, but it would not satisfy the needs of a regional program officer in tracking results from the SDPs in the program network.

In contrast, there are techniques that monitor achievements of a program but provide relatively little insight into strengths and weaknesses. One example would be the routine reporting of service statistics (e.g., number of new acceptors, number of clinic visits, number of couple–years of protection, etc.) collected at the level of the SDPs and aggregated at a central level. This type of information is valuable in tracking trends over time, yet alone does not indicate why the program is or isn't achieving the desired results.

Program administrators and donor agencies are generally interested in both types of monitoring. Donor agencies almost always want the quantitative data on "results," but are increasingly interested in knowing that the program has some means of obtaining data, often qualitative in nature (e.g., focus groups, in–depth interviews, observation checklists, etc.), that will be used directly for program improvement. To this end program monitoring often consists of a combination of evaluation activities that collectively provide information on the program as a whole.

IDENTIFYING THE COMPONENTS TO BE MONITORED

The decision as to which components of a program to monitor depends in part on the primary purpose of the evaluation: to improve the program, to track results, or both (see Figure III–1). For the program manager, it is not a question of "one or the other;" he/she will need both. By contrast, donor agencies are often more interested in tracking program–level results

Figure III–1

Purpose of Monitoring and Components of Interest

Purpose	Components
Improving the Program	*Functional outputs* Number and quality of activities conducted in different areas of management/supervision, training, commodities logistics, IEC, record keeping.
	Service outputs Access to services, quality of care, and program image.
Tracking the Results	*Service utilization* Results produced at the program level (e.g., number of acceptors, number of visits, CYP, etc.).
	Outcomes Intermediate or long–term changes at the population level (e.g., contraceptive prevalence, median interval between births, TFR).

(although in general they strongly encourage implementing agencies to carry out evaluation activities intended to identify areas in need of improvement).

Evaluation of the different service operations ("functional areas") is particularly useful when done early enough in the implementation process to allow for mid–course corrections. Service utilization is generally tracked continuously over the life of the project. Outcomes, by contrast, are generally measured at (two or more) intervals to measure change over time.

Certain points warrant mention:

- Program monitoring ideally employs both quantitative and qualitative techniques.
- It is not practical to attempt a detailed evaluation of ALL aspects of the program. Rather, it is important to prioritize those points for which the information will be most useful to the organization and crucial to the success of the program.
- The use of a conceptual framework to define the pathways to achieving the desired results is equally applicable to projects as to national programs (see box on the following page).
- Evaluation activities are generally staggered. Some may be done routinely (e.g., collection and reporting of service statistics), others on a periodic basis (e.g., simulated or "mystery" client surveys to assess quality of care), and others as a one–time exercise (e.g., analysis of cost per CYP for different contraceptive methods).
- The key management staff should take the lead role in deciding what aspects of the program to monitor, not the evaluation specialist.[7] In this way, evaluation truly serves the needs of the organization.

To arrive at a decision regarding specific aspects of the program to evaluate, it is useful to identify the range of possible topics and then prioritize them. Figure II–6 (on page 18 of the previous chapter) provides a useful framework for considering options. It is replicated in Figure III–2 with the spaces left blank, which an organization could use to identify and prioritize evaluation questions.

[7] It is important, however, for the evaluation specialist to be closely involved in these discussions, to provide information that may influence the decision (e.g., the approximate costs and time required for different data collection activities, alternative sources of information, the biases inherent in different methods of data collection, and so forth).

The task then is to identify the most pressing evaluation questions for the program under study. For example, almost all managers will want regular feedback on "results" that reflect performance at the facility level (e.g., number of new acceptors, number of CYP). In addition, there are often specific concerns to be addressed (e.g., the need to monitor quality). Managers will also want to know how productive their staff is in producing outputs such as client visits or CYPs; for example, how many clients receive services per day? What was the labor cost of producing CYPs in the previous year? How did the cost per CYP vary by type and location of clinic?

DEFINING RELEVANT INDICATORS

The Purpose of Indicators

Indicators are variables that measure the different aspects of a given program: the inputs, processes, outputs, and outcomes.

An indicator can be assigned a numeric value (a percentage, a mean value, a ranking, an absolute number) or a yes/no score (e.g., "presence" versus "absence"). In some cases, the value carries a widely shared interpretation (e.g., a contraceptive prevalence rate of 75 is "high"). In other cases, the value (number) is most useful in a relative sense, in

Conceptual Frameworks are Useful in Evaluating Projects and Programs

Chapter II described in detail the value of a conceptual framework in designing an evaluation of a national program. However, it is equally useful to have a "road map" that describes the pathways to achieving desired results for smaller–scale projects. Although the expected change is generally expressed in terms of a practice or behavior, that may not be the case in some programs, as shown in the example below, on youth and AIDS prevention.

In this example the implicit assumption is that knowledge obtained at a younger age will in fact influence behavior among young people as they become at risk for HIV infection, but the scope of the evaluation (and the period for the project) may not allow for the testing of this assumption. Rather, as reflected in the framework below, the project is designed to increase knowledge of HIV transmission, and the evaluation would focus on this result. The framework is useful in clarifying the pathways to change. Each intermediate result can in turn be monitored to assure that the project is being implemented according to design.

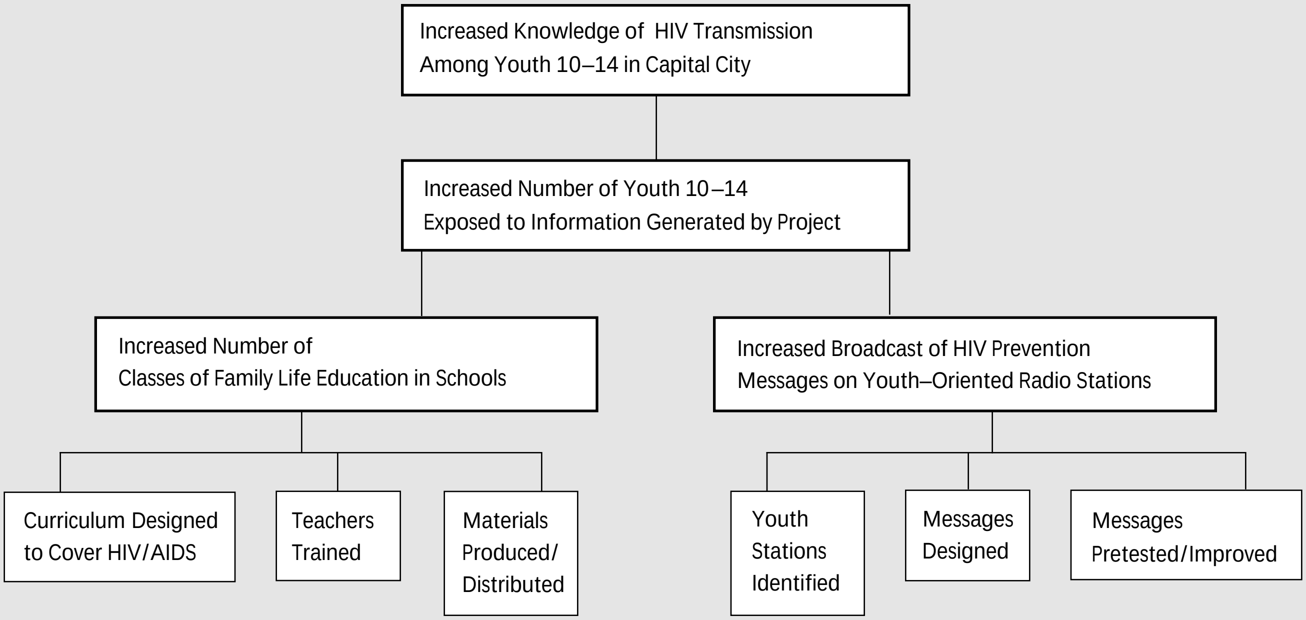

Increased Knowledge of HIV Transmission Among Youth 10–14 in Capital City
Increased Number of Youth 10–14 Exposed to Information Generated by Project
Increased Number of Classes of Family Life Education in Schools
Increased Broadcast of HIV Prevention Messages on Youth–Oriented Radio Stations
Curriculum Designed to Cover HIV/AIDS
Teachers Trained
Materials Produced/Distributed
Youth Stations Identified
Messages Designed
Messages Pretested/Improved

Figure III–2

Framework For Identifying and Prioritizing Aspects of Program Performance to Monitor

Program Component	Quantity of Component	Quality of Component	Cost of Component
Inputs			
Outputs Functional areas: ➤ Management ➤ Training ➤ Commodities/ Logistics ➤ IEC ➤ Research/ Evaluation			
Outputs Service outputs	Access:	Quality:	
Outputs Service utilization			
Outcome Intermediate outcomes			
Outcome Long–term outcomes			

comparison to similar programs or the same program at an earlier date (e.g., trends in CYP).

At the risk of oversimplification, program monitoring consists of measuring how well the program is doing in one or more of the "boxes" of the conceptual framework (see Figures II–3 and II–4). The framework illustrates how the program should theoretically work in achieving the desired results at the program and population levels. Program monitoring quantifies what actually occurs at each level (of inputs, processes, outputs, and outcomes). Whereas some might consider the conceptual framework only an academic exercise, in fact it is extremely practical, given that it identifies the areas for which the evaluator may want to select indicators.

A menu of possible indicators for evaluating family planning programs is provided in the *Handbook of Indicators for Evaluating Family Planning Programs* (Bertrand et al., 1994). As family planning programs expand to include other aspects of reproductive health, the potential number of other relevant indicators expands (see Bertrand and Tsui, 1995, which describes indicators for the areas of safe pregnancy, STD/AIDS, women's nutrition, breast-feeding, and adolescent reproductive health services).

For a given evaluation, one should prioritize indicators based on specific program objectives and select a manageable set of indicators to meet the particular needs of the situation. In short, it is essential to identify the key question(s) to address in the evaluation and to select the indicators accordingly.

Characteristics of Good Indicators

Good indicators share some important characteris tics. Perhaps the single most important is validity. Does the indictor measure what it is supposed to measure? For example, survey questions on ideal family size are not generally thought to be valid measures of fertility demand due to reporting biases. On the other hand, the stated intention to have more children is thought to be a valid indicator of demand since it tends to be less influenced by reporting error. Sales of socially marketed contraceptives may be a valid indicator of marketing success, but an invalid indicator of the level of contraceptive use, if users are substituting the social marketing methods for other methods.

Another dimension of validity is that the indicator should have a close or at least defensible connection to the intervention. For example, the maternal mortality rate in general is not a valid measure of the impact of a family planning program on women's health. While family planning programs certainly contribute to reducing maternal mortality, there are numerous other factors (prenatal care, referral system, accessibility of hospital care, transport) that also influence that rate. A more valid indicator of the impact of family planning on women's health may be a measure of births averted among women known to be in high risk categories.

Reliability is another desirable characteristic of a good indicator. Reliability refers to the degree of random measurement error in an indicator. Measurement error may arise from sampling error, non–sampling error, or subject measurement of the indicator. For example, due to sampling error, a national survey such as the DHS will not provide reliable estimates of contraceptive practice for small areas because of large sampling errors. Service statistics may yield more reliable measures of contraceptive practice in these areas (though they may not be valid if the population in the area obtains contraceptives elsewhere). Sample surveys generally provide unreliable estimates of abortion due to response bias, in this case the reluctance of respondents to report abortions. In addition, indicators that rely on subjective judgments, for example the quality of program leadership, may be unreliable, as different evaluators may use varying standards to measure that characteristic.

An indicator should be defined in clear, precise terms. The indicator must be operationally defined so that others will know precisely what is being measured. For example, what is a family planning acceptor? How is provider competency defined? Is knowledge of a family planning method based on spontaneous or prompted recall? The definition of the indicator should also specify the population among which the indicator is measured: all women or married women? Women aged 15–44 or 15–49? All family planning users or just users of modern methods? All women in the country or only those in the project area? Finally, the meaning of demographic measures such as rates and ratios should be clearly specified.

Similarly, it is preferable to choose indicators that are comparable across different population groups and program approaches. All things being equal, a contraceptive prevalence rate based on women 15–49 is preferable to one based on women 15–44, because it will be more comparable to rates from other programs. A program with different service delivery approaches (e.g., clinic, CBD, and social marketing) should try to identify at least one or two indicators of service utilization that are appropriate to all three modalities, such as CYP, so that results can be compared across service delivery approaches. Where possible, indicators should reflect performance relative to some standard or "denominator." For example, it may be of interest to know the number of CYP for a specific method from the commercial sector, however, it is perhaps more useful to know the "percentage of users of a specific method who obtain it from the commercial sector" (for re–supply methods).

Indicators should be non–directional in nature. They describe the situation at a given point in time. If, for example, a program is trying to assess its progress in decreasing stockouts, the appropriate indicator is the "percentage of SDPs that encountered a stock–out during the past 12 months" (which could be tracked over time), not "a decrease in the percentage of SDPs that had a stock–out."

Indicators should be collected on a timely basis. The indicator should provide a measurement for a recent period or at least for the period during which the intervention occurred; also, it should be available at appropriate intervals. Population–based indicators are rarely available annually and often refer to a period of several years before the survey (for example, DHS estimates of fertility typically refer to the three–year period that precedes the survey). While routine program–based data such as service statistics would seem to be a good source for current data, there is often considerable delay in their availability. Finally, some program–based data are now being collected periodically through such methods as Situation Analysis or the DHS Service Availability Module. These instruments provide timely, but not continuous, measures of program functioning.

Factors that Affect the Selection of Indicators

In an ideal world, the evaluator would systematically identify the indicators judged to be most useful for a given evaluation and proceed to collect or acquire the needed data. However, in the field setting where time, human, and financial resources are in short supply, others factors intervene in the selection of indicators. The following are common factors that enter into the decision.

- Availability of data needed to measure the indicator

  Example: To assess the effects of family planning programs on fertility and health outcomes worldwide, it would be extremely useful to have data for all countries on the sources of funding (donor agencies, local taxes, client fees) for family planning and on costs of providing family planning services. However, such data do not currently exist in readily accessible form. Moreover, it is unclear whether all governments would be willing to open their financial records to outside evaluators for the purpose of collecting this information.

- Amount of time allotted to the evaluation

  Example: Program managers might like to know whether their new approach to counseling NORPLANT® clients results in longer continuation rates. However, if the evaluation of the counseling program is limited by the life of the project (part of which has presumably elapsed), it is impossible to ascertain the long–term effects of this counseling.

- Financial support available for evaluation

  Example: Many IEC directors would like to know the percentage of the target population reached by a given campaign and the reaction of the public to the messages. However, they may not have the resources for conducting a population–based survey. There may be a trade–off between cost on the one hand and validity, reliability, and timeliness on the other hand.

- Donor agency requirements

  Example: The indicator "couple–years of protection" has become the most widely used measure of service utilization in USAID–funded programs, because USAID (as well as IPPF) requires recipient agencies to report this result.

Use of Multiple Indicators

For well–established indicators such as the Total Fertility Rate (TFR) and the Contraceptive Prevalence Rate (CPR), single indicators are usually sufficient. However, there are instances when it may be advisable to use two or more indicators to measure a given result. One such situation is where data quality is suspect; a given result is more credible if the same trend can be demonstrated across two or more indicators.

Secondly, when new program indicators are being introduced, it is useful to have alternative indicators for a given category of result (e.g., to measure the quality of the client–provider interaction). This provides the evaluator with a back–up plan in the event that the data from one source do not materialize or are judged invalid (e.g., respondents misunderstood the question). Finally, the indicators in a given functional area often measure a chain of events, and the use of multiple indicators may be important to developing an understanding of the dynamics along the chain. For example, an IEC program (1) generates a certain number of messages via a certain number of channels and (2) offers counseling to potential and actual clients who seek services in the expectation that members of the target population will (3) hear messages about family planning, (4) understand the main messages, (5) react positively to the messages, (6) discuss the messages with others, (7) develop a favorable predisposition toward the behavior, such as contraceptive use, (8) become an acceptor, and (9) continue the practice. At each step in the process the percent following this chain can potentially decrease; to the evaluator it is indispensable to identify the pattern of this response on the part of the target population.

In sum, the selection of indicators is based on the purpose of the evaluation, to learn more about a specific functional area, program output or outcome, and in some cases to meet donor agency requirements. The selection of indicators is dictated by the specific needs and interests of those undertaking the evaluation. Different types of evaluations may be staggered over the five–year (or longer) lifetime of a project.

Operationalizing the Indicators

"Operationalizing the indicators" means identifying how a given behavior or concept will be measured. In the best case scenario, an indicator is conceptually clear and lends itself to easy, unequivocal measurement. An example would be the number of persons trained in a given year, by category of personnel (physician, nurse, commodities/logistics specialist, etc.).

Unfortunately, very few of the indicators are so simple and straightforward. Rather, even after the evaluator has identified the indicators to be used, he/she tends to be faced with one or more of the following problems in operationalizing them.

- The measurement of an indicator requires subjective judgment.

  Many agree that one of the single most important factors in the successful family planning programs currently in existence is the "quality of program leadership." To use this indicator, it is imperative to define the characteristics that constitute "leadership," but the final assessment remains subjective.

  Similarly, indicators requiring a judgment of "presence" or "absence" may be difficult, where it is necessary to establish "how much" constitutes "presence." For example, with regard to the indicator "absence of unwarranted restrictions on users," one could have a model program in this sense with one small exception. Should that exception count for enough to alter one's assessment?

  If faced with the problem of subjective judgment calls, the evaluator must first decide whether to retain the indicator. If so, then he/she should clarify the criteria used in arriving at the final score or assessment.

- The rules of measurement are clear but local applications differ from the recommended approach.

  For example, the *Handbook of Indicators for Family Planning Program Evaluatio*n (Bertrand et al., 1994) recommends defining "number of acceptors new to the institution" as "new only once." That is, if a person drops out of the institution's program for several years and eventually returns, then she would NOT be "new to the institution." However, since some organizations don't retain records after a five year period, it may not be feasible to adhere to the recommended definition. In such cases, the evaluator should be very clear how the measurement used differs from standard or recommended practice.

- The indicator is conceptually clear but the "yardstick" for measuring it is not.

  At first blush, the "cost of one month's supply of contraceptives as a percentage of monthly wages" appears to be clear and measurable. However, as one applies the indicator, certain questions may arise. For example, different contraceptive methods have different costs. What cost should be included: the average cost of all methods available? The average cost weighted by the proportion using the different methods? Moreover, what numbers should be used if the cost of methods varies over the course of the year? What number should be used in the denominator if average monthly wage figures are outdated?

  To the extent that evaluators have access to reports by others, it is useful to review how fellow researchers and evaluators have handled similar situations. In the absence of such information, the cardinal rule bears repeating:

document whatever decisions are made in operationalizing the indicators for a specific evaluation.

For every indicator that might be used in program monitoring, it is necessary to identify one or more sources of data from which to obtain the information, as described in the following section.

IDENTIFYING SOURCES OF DATA

The evaluation of a national family planning program usually entails both population–based and program–based data. Whereas multiple sources of data exist (see Figure III–3), most family planning evaluation is limited to a few sources of data. In designing an evaluation, it is essential to inventory the data that already exist and to identify additional data collection that is necessary to provide answers to the evaluation questions under study.

Which comes first: identifying the indicators to use or selecting the sources of data for measuring them? The two processes are closely intertwined; the choice of indicators is often dictated by the availability of existing data or the feasibility of collecting additional information at minimal cost to the program (e.g., routine service statistics). However, some types of evaluation can NOT be carried out using existing data, such as assessments of quality of care that require special studies. In such cases, one attempts to outline the indicators of interest, then to identify the data collection required to obtain the information.

Although there are potentially a large number of data sources for evaluating family planning programs, the vast majority of program evaluation is based on the following sources of data:

Component to Measure	Source(s) of Data
Outputs (program–based)	Program records, especially service statistics Facility surveys Data on the commercial sector Special studies
Outcomes (population–based)	DHS–type household surveys

Program–based Data

There are many different types of information that constitute program–based data. These can be summarized in four main categories.

Program Records and Service Statistics for Public Sector and NGO Providers

Program records refer to all types of information generated by one or more divisions of the program and kept on file at a central or regional office. Examples useful to monitoring programs (specifically, functional outputs) include: the number of supervisory visits made to CBD workers, the number of persons trained per year by type of personnel, the number of different communication products produced in a given year, the quantity of each communication product distributed, and so forth.

Service statistics are a sub–category of program data. They include any type of information routinely collected and reported with regard to the utilization of a service. Common indicators in family planning programs include:

- number of new acceptors;
- number of visits to the SDP;
- couple–years of protection; and
- user characteristics.

Generally, these data are collected at each SDP, then aggregated in a central office to monitor trends over time and by sub–unit within the system.

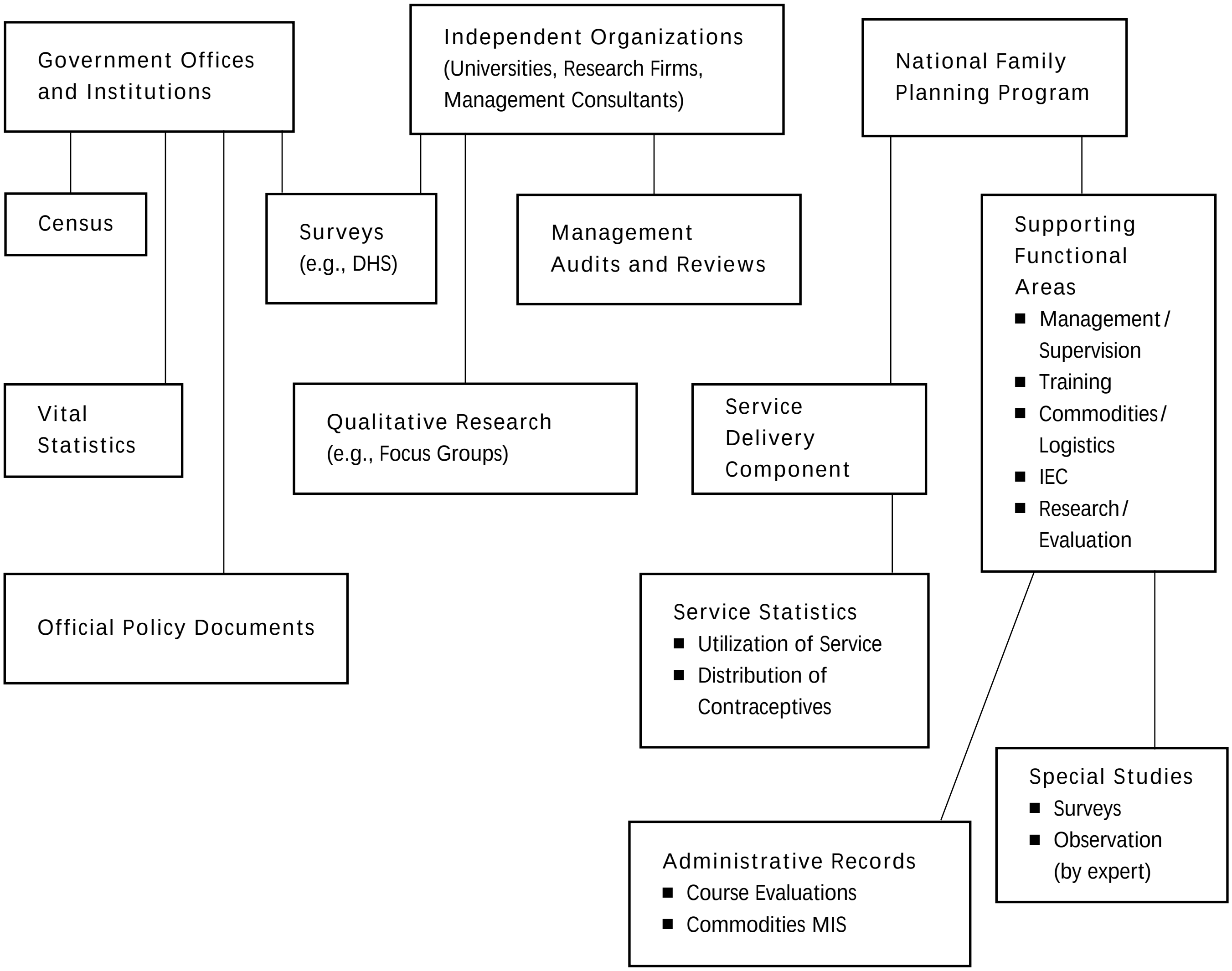
Facility Surveys

The prime objective of facility surveys is to describe the availability, functioning, and quality of health and family planning activities. This information can be obtained by interviewing informed respondents or by visiting the facility and observing its operations. Facility data are also required for studies that link information from the program level (e.g., quality of care) with outcomes at the population level (e.g., contraceptive prevalence) for example, see Mensch et al. (1994).

There are two main types of facility surveys used in connection with USAID–funded activities. The first, known as the "Situation Analysis," was developed in the context of the Africa Operations Research/Technical Assistance (OR/TA) project (Fisher et al., 1992) and has been replicated in numerous countries around the world. The second

Figure III–3

Sources and Type of Data for Family Planning Program Evaluation


Government Offices and Institutions
Independent Organizations (Universities, Research Firms, Management Consultants)
National Family Planning Program
Census
Surveys (e.g., DHS)
Management Audits and Reviews
Supporting Functional Areas
■ Management/ Supervision
■ Training
■ Commodities/ Logistics
■ IEC
■ Research/ Evaluation
Vital Statistics
Qualitative Research (e.g., Focus Groups)
Service Delivery Component
Official Policy Documents
Service Statistics
■ Utilization of Service
■ Distribution of Contraceptives
Special Studies
■ Surveys
■ Observation (by expert)
Administrative Records
■ Course Evaluations
■ Commodities MIS

is the Service Availability Module (SAM), developed and implemented under the DHS program.

The two types of surveys differ (1) in the data collection instruments used and (2) in the population of SDPs that they describe. For Situation Analysis, the instruments for collecting the data include a series of modules (e.g., inventory of the SDP, observation of provider–client interaction, exit interviews with clients, interviews with service providers). Descriptive results provide program administrators at the central level with important insights into the strengths and weaknesses of the program throughout the country (or geographic area under study). The Situation Analysis is also used to obtain data on the quality of care in family planning programs.

The SAM is a community–based survey, conducted in connection with the household level survey in the DHS (in selected countries). For every sampling cluster used in the study, key informants provide a list of existing health/family planning facilities. Teams are then dispatched to collect data at the nearest (1) hospital, (2) clinic, (3) health center, (4) pharmacy, and (5) private doctor within a 30–kilometer radius of the center of the cluster.[8] The information collected at each location covers the governing structure (private versus public), number and type of staff, infrastructure (equipment, type of construction materials), types of services provided, types of contraceptive methods available, etc. This information is potentially useful for assessing the availability and adequacy of family planning services for a given population, and it can be particularly important in linking changes (improvements) in the family planning supply environment to changes in prevalence over time.

The difference in the sampling used for the two types of surveys is as follows: the Situation Analysis is based on a random sample of SDPs in a country (which may be disproportionately located in urban areas), whereas the SAM data are collected with respect to a random sample of women in the country.[9] Thus, the Situation Analysis measures the average SDP, whereas the SAM measures the services available to the average women in a given country.

To date, facility surveys have been greatly under–utilized for the purposes of evaluating family planning programs. However, their potential utility in assessing impact (in connection with the individual interviews) is discussed in Chapter IV.

Cost studies can also be conducted at facilities. These provide information on the cost of providing different services including acceptor and follow–up visits by method. Using information on visit patterns and the number of CYP by method, these data can be aggregated to determine the cost per CYP. In addition, the cost of expanding acceptor and follow–up visits as well as the number of CYP can be calculated taking into consideration the amount of under–utilized capacity.

Information on Family Planning Provision in the Commercial Sector

Data sources on the commercial sector are generally not part of the program statistics, which are usually maintained by the public sector and large NGO providers. In fact, there is often no single source on family planning service statistics in the commercial sector. In part, this is due to the absence of a central body who would be responsible for this data and in part, because commercial providers are competing with one another and may be reluctant to share information on the volume and quality of their services.

Information on family planning activities in the private, commercial sector have been measured in some SAMs and Situation Analysis Studies as well as in surveys focused exclusively on private providers. Additionally, some data sources unique to this sector are available to measure the availability of methods and services in the sector. These include:

- data on contraceptive shipments to distributors and wholesalers;
- sales at the retail level;
- audits of retail outlets; and
- reports from detail men on the availability of family planning services and methods from private doctors and clinics.

Providers in the commercial sector and supporters of social marketing and private sector programs are also concerned with the quality of services provided. Some of the techniques used to assess the quality of services are: (1) mystery shopper studies to determine whether retailers are promoting social marketing products and to assess the quality of the information provided; (2) consumer intercepts to assess consumer satisfaction with the services received; and (3) population–based surveys.

Special Studies

Special studies are generally conducted to respond to a specific need. They may employ quantitative or qualitative research methods. The list of possible special studies is long; illustrative examples include the following:

- a follow–up of sterilization clients to determine their level of satisfaction with the procedure;

[8] It is preferable for the data collection team actually to visit each site; however in earlier days of the DHS or where funds were not available, key informants provided this information in some of the studies.

[9] However, with appropriate weighting, results from the SAM can be presented for a population equivalent to that usually described by the Situation Analysis Study.

- focus groups among adolescents attending a given program to assess whether it responds to their interests and needs;
- a management audit of program documents; and
- mapping of a community to show where eligible couples live and what method of contraception is used.

Population–based Data

The primary tool for collecting population–based data for family planning programs is the DHS–type survey. Following in the tradition of the World Fertility Survey (WFS) and the Contraceptive Prevalence Survey (CPS), the Demographic and Health Survey (DHS) is generally conducted among a representative sample of women of reproductive age in a given country. In recent years, a number of DHS surveys have also included a sample of men (either an independent sample of men or a sample of husbands of women interviewed for the DHS). The DHS core questionnaire consisting of some 250 questions provides detailed information on fertility and family planning, in addition to information on maternal and child health, health services utilization, and related topics (Robey et al., 1992).

We use the term “DHS–type surveys” to underscore that there are other surveys similar to the DHS in content and type of population studied. The Centers for Disease Control and Prevention (CDC) have conducted a number of Reproductive Health Surveys in selected countries of Latin America and other regions of the world. Similarly, certain countries have conducted their own national–level surveys on fertility and related issues.

Figure III–4

New Acceptors by Method and Year: 1980–1993

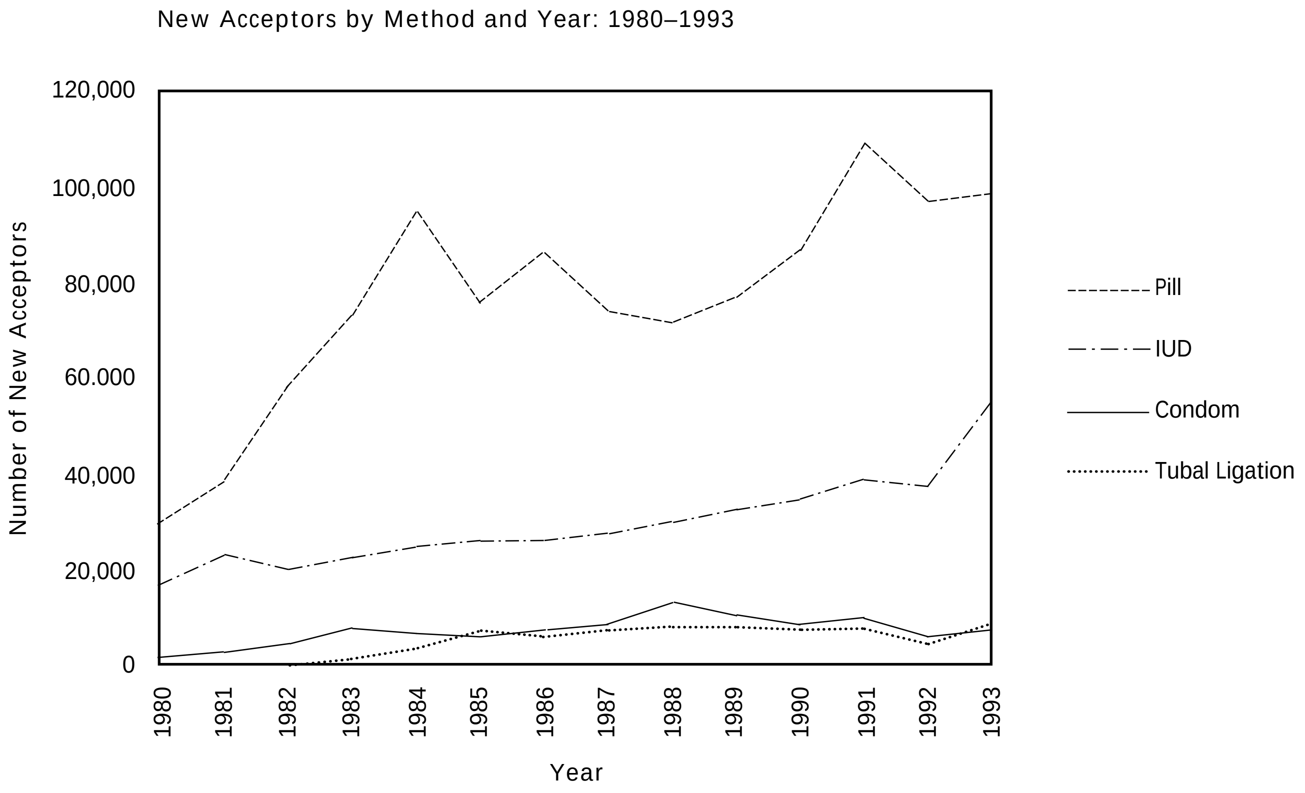
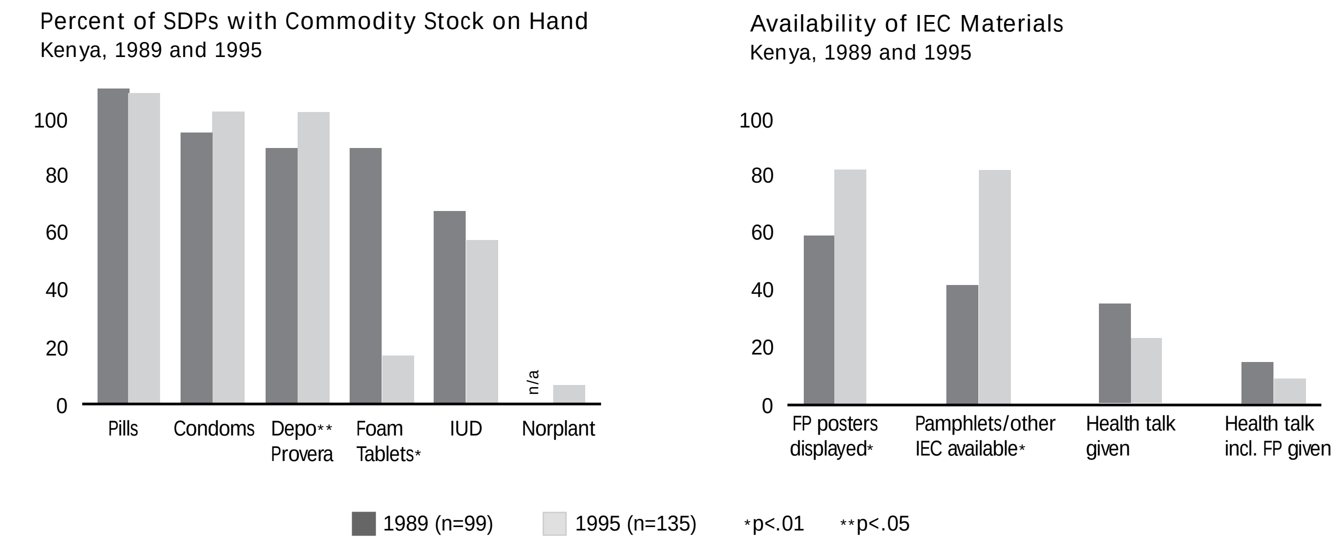


Figure III–5

Changes in Service Outputs at Two Times, as Measured by Situation Analysis



Source: Miller et al., 1996. "A Comparison of the 1995 and 1989 Kenya Situation Analysis Study Findings," unpublished manuscript.

Figure III–6

Method Mix Based on National Surveys

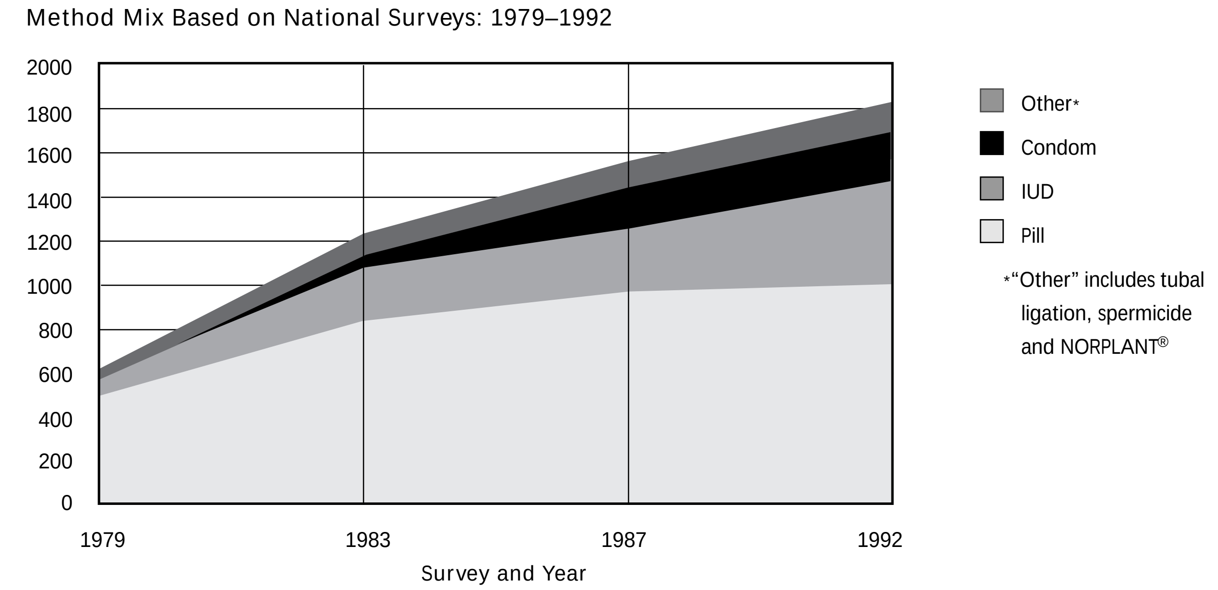
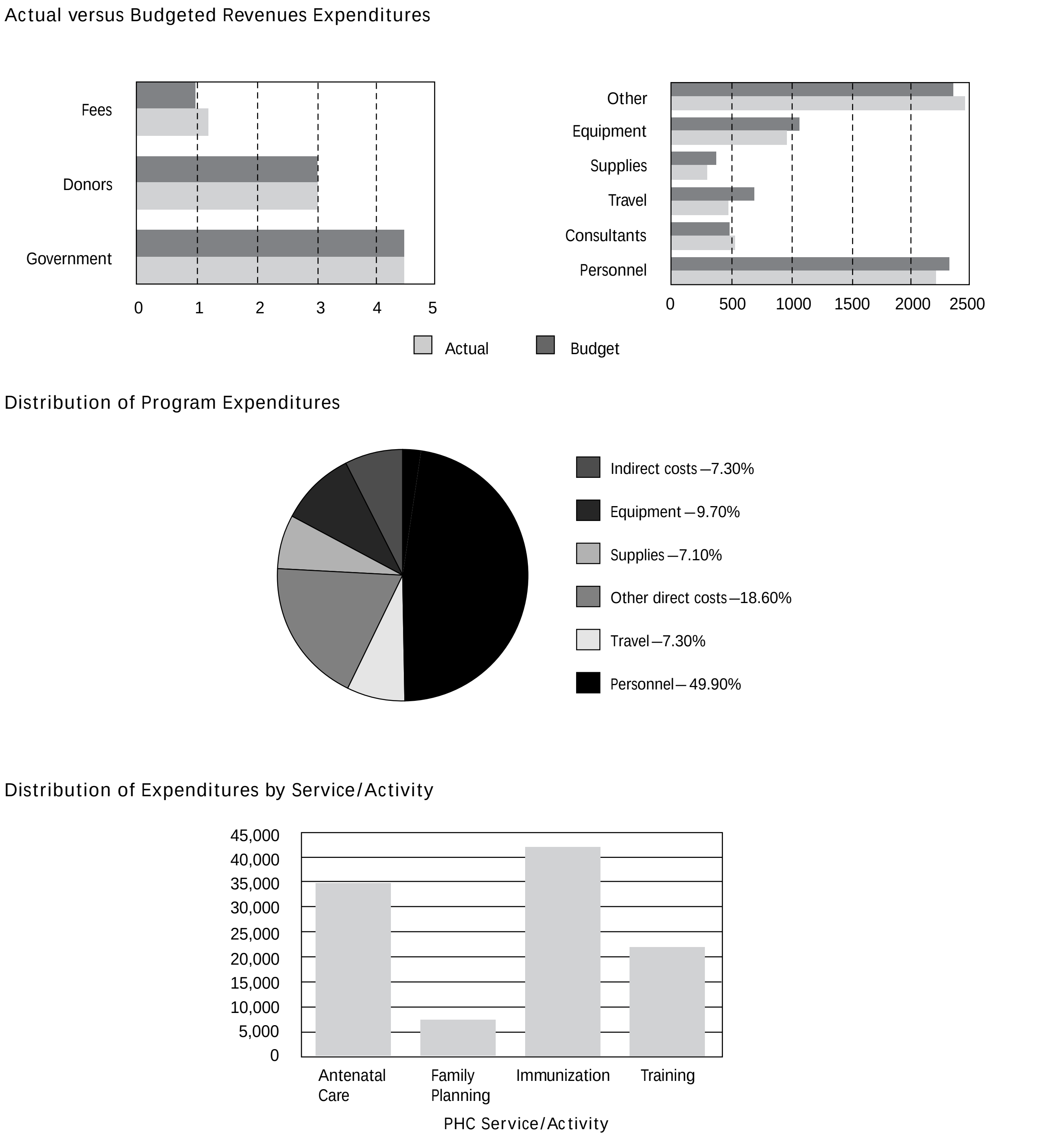


Figure III–7

Illustrative Results: Cost Data From Program Records



Actual versus Budgeted Revenues Expenditures

Distribution of Program Expenditures

Distribution of Expenditures by Service/Activity

Source: Reynolds, J. 1993. Cost Analysis: *Primary Health Care Management Advancement Programme, Module 8 Users Guide*, Washington, DC: Aga Khan Foundation and University Research Corporation, page 9.

DESIGNING A FORMAT FOR PRESENTATION OF RESULTS

Frequently, evaluators and other researchers collect far more data than they need. One way of avoiding this is to map out how the information will be processed and presented well in advance of the actual data collection. This approach allows the evaluator to see what the report will look like and to assess whether the amount of detail is appropriate to the situation. Conversely, this step may also bring to light the lack of key information which should be included in the data collection.

Researchers often refer to this step as "designing the dummy tables." That is, one produces a series of tables and figures that indicate the exact variables to be presented in each and the type of data to be included (absolute numbers, percentages, means, medians, etc.), but none of the actual data (which are not yet available or processed).

Data may be presented in tabular or graphic form. An advantage of presenting the results in tables is that one can include precise values for a fairly large number of indicators. However, the "sea of numbers" may discourage some readers from delving in to find the trends. By contrast, graphs tend to highlight the trends in the data (thus increasing comprehension), but may not allow for presentation of specific values. The basic techniques for presenting results are similar, whatever the source of data. In the figures that follow, we give examples from common sources of data of typical indicators one might track:

- routinely collected service statistics: number of new users by type of method (Figure III–4);
- facility–based data (Situation Analysis): percent of SDPs with commodity stock on hand, availability of IEC materials (Figure III–5);
- population–based measures of outcome (DHS): changes in method mix over time (Figure III–6); and

Figure III–8

Example: Overview of the Methodological Approach to be Used in Monitoring a Specific Program

Components to be Monitored	Source of Data	Indicators	Frequency of Reporting	Organization Responsible
Access to Services	Program records	Number of SDPs offering contra–ception in a defined geographical area	Quarterly	MOH and private FP association
Quality of Care	Facility–based survey (e.g. Situ-ation Analysis)	Illustrative: ■ Number of methods in stock ■ Provider gives accurate, relevant information on method accepted	Occasional	Organization analyzing Situation Analysis data
Service Utilization	Service statistics	Number of acceptors, CYP	Quarterly	MOH and private FP association
Costs	Financial data and other program records	Cost of acceptor, follow–up, and discontinuation visits by method	One–time study	Organization responsible for study

- financial records of the program: expenditures in comparison to amounts budgeted, expenditures by type and by service activity (Figure III–7).

In designing this set of dummy tables and figures, the evaluator should include all the information that he/she intends to present. From this exercise, one can assess quickly whether the evaluation results will be adequately comprehensive without overwhelming the potential user.

SUMMARIZING THE METHODOLOGICAL APPROACH

There is no set rule about the form in which the issues outlined in this chapter need to be summarized in the evaluation plan. However, it is important for the document to present a global view of the elements covered in this chapter:

- primary purpose(s) of monitoring;
- components to be monitored (based on the "road map" of the conceptual framework);
- indicators for each data collection exercise;
- source(s) of data for each data collection exercise (with an indication of which exist already and which will need to be collected); and
- format for presenting the results.

Chapter IV

Methodological Approach: Impact Assessment

- A Friendly Caveat
- Overview of the Issues
- Criteria Guiding the Choice of Methodological Approach
- Preferred Approaches
- Alternative Approaches
- Summary

Chapter IV

Methodological Approach: Impact Assessment

A FRIENDLY CAVEAT

Chapters I to III of this document dealt with issues and concepts that are likely to be at least partially familiar to most readers. The material was presented in a step–by–step format so as to further clarify the program evaluation process and make the tools of program monitoring more readily comprehensible to all readers, whatever their statistical and research background.

By contrast, Chapter IV addresses a series of issues that are more methodologically complex. The goal of the chapter is to promote a general understanding of (1) the relative strengths and weaknesses of various approaches and methods for measuring program impact, (2) the rationale for The EVALUATION Project's endorsing some methods and not others, and (3) the general requirements for using these methods. The authors have attempted to present the material in language that communicates the basic methodological ideas to readers with limited background in statistics and research, while at the same time providing sufficient technical information to readers with more advanced training such that the relative technical merits of the alternative approaches can be appreciated. However, the presentation in this chapter falls short of providing the same step–by–step guidance found in the first three chapters, and the computational details are not presented. References are provided to other publications that give the necessary technical details for the application of the methods discussed in this chapter.

OVERVIEW OF THE ISSUES

Increasingly, attention is being focused on the "bottom line" of family planning and health program efforts; that is, on the impact of programs or interventions. As noted in Chapter II, measuring impact entails more than merely monitoring changes in outcome indicators; it requires plausible evidence that an observed change in outcome indicators is attributable to a given program or intervention. In other words, it requires evidence that the program or intervention being evaluated has caused the change to occur.

A number of different methods or approaches are available for measuring program impact. These differ in important ways, for example, in terms of the types of outcome measures used, the number and types of assumptions required, the strength of conclusions regarding program impact reached, and operational and data requirements. This chapter provides an overview of what The EVALUATION Project views as the more robust methods available for measuring family planning program impact, the objective being to provide guidance in choosing among the available options.

The material in this chapter was prepared with the program/evaluation officer or evaluation researcher responsible for preparing the evaluation plan for a new program or program cycle in mind. Decisions as to how program impact should be measured are best made in the design or planning phase of programs or new program cycles, because it is only at this stage that the full range of options is available. Once a program or program cycle is in progress, the opportunity to use some of the stronger research designs and methods has been lost, and it becomes necessary to resort to weaker options.

In contrast to monitoring trends in program statistics or conducting descriptive studies for program improvement (e.g., client satisfaction surveys), assessing impact in the strict sense of the term requires statistical methods and analytic skills for which many program administrators, donor agency staff, and host country evaluation specialists may not have been trained. In this chapter, we present a description of the different methods intended for readers with limited statistical background, which nonetheless would allow them to judge the merits of using a given technique in their own setting (even if others were to carry out the actual analysis). These general descriptions are followed by a discussion of key methodological issues and a summary of strengths and weaknesses, intended to provide statisticians and demographers with further insights into the methodological implications of different method choices.

CRITERIA GUIDING THE CHOICE OF METHODOLOGICAL APPROACH

Given that the primary purpose of this chapter is to assist readers in choosing among alternative methods or approaches for measuring program impact, it is useful to begin by specifying some criteria for making such choices. In assessing the various methods described in this chapter, the evaluator should take the following criteria into consideration:

- Exposure to threats to validity[10]

  The single most important criterion in assessing a method is the validity of its estimates of program impact. While all methods are vulnerable to some threats to validity, they vary considerably in terms of the number and types of threats (i.e., confounding factors) to which they are subject.

- Required assumptions

  Assumptions are required of all methods. Some methods require strong assumptions that are rarely valid in actual practice, while others require assumptions that are weaker and more likely to be valid. Methods requiring fewer and less stringent assumptions are to be preferred, other things being equal.

- Ability to isolate program effects

  Ideally, measures of program impact will include only results that are directly attributable to the program. In most settings, factors such as the forces of socioeconomic development, multiple social programs, changing demographic structure, and the presence of non–program family planning activities complicate attempts to measure program impact. Methods differ in the extent to which the evaluator can isolate program effects from the influences of other factors; and methods that are more efficient in doing so are preferred.

- Cost

  This refers to the costs of data collection and analysis. Other things being equal, less costly methods are preferred. In most cases, however, methods differ on other criteria as well as cost, and thus cost–benefit decisions must be made.

- Data requirements

  Methods vary considerably in data requirements. Aside from differences in the volume of data needed, some methods require data that are more difficult to collect and/or are more vulnerable to measurement error than other methods, and thus increase the risk that measurement error may obscure actual program effects or exaggerate the magnitude of impact actually achieved.

- Insights into causal pathways

  Methods vary considerably in the amount of information they provide as to how program inputs are transformed into outputs and outcomes as part of the impact measurement process. Although not required for the measurement of impact, such information provides useful insights as to how programs might be improved in subsequent program cycles.

- Types of outcome indicators used

  Some methods have been designed specifically for the measurement of certain types of outcomes. For example, a number of methods have been developed specifically to measure the fertility impact of family planning programs that are not readily adaptable to measuring impact of other outcomes (e.g., health outcomes). Other methods are more versatile and may be used to measure other types of program results, as well

[10] The term “validity” as used here means the extent to which measurements of program impact from a given study design constitute unbiased and unconfounded measures of actual program impact. In lay terms validity refers to the fact that one is actually measuring the phenomena one intends to measure.

as results at different levels. Thus, the types of outcomes of interest in a particular evaluation effort will in part dictate choice of method.

- Degree of program control required

  While certain types of research designs (e.g., randomized experiments and to a lesser extent quasi–experimental studies) provide the strongest evidence of program impact, they also require more controlled conditions in terms of the manner in which the program being evaluated and other interventions being implemented during a given study period are undertaken. Other approaches falling under the heading of "non–experimental" studies do not require that programs be implemented in specific ways in order to provide valid measures of program impact, but generally require larger quantities of data and more complicated analysis in order to produce valid findings. The degree to which program activities can realistically be controlled by implementing agencies so as to facilitate impact measurement is thus an important factor in choice of study design.

- Technical/statistical skills and resources required

  While all of the methods considered require basic research and statistical knowledge and skills, some of the methods and approaches require relatively advanced skills and in some cases specialized computer software.

Thus, a fairly large number of factors need to be considered in deciding upon an approach for measuring program impact. To facilitate the choice of an appropriate method, we have classified methods for impact measurement into two categories, reflecting what the authors perceive to be their overall strength based upon the criteria outlined above:

- Preferred approaches

  Methods falling into this category are viewed as being the strongest designs (for reasons indicated below) and are recommended as a first choice wherever possible.

- Alternative approaches

  Where the use of preferred methods is not possible, several useful alternative approaches are available. These methods are based upon less rigorous designs and generally produce less compelling results than the preferred methods, but are capable of producing defensible estimates of program impact under certain circumstances.

The specific methods considered in this chapter and their classification into the two groups defined above are shown in Figure IV–1.

Figure IV–1

Classification of Methods on the Basis of their Overall Utility as "Stand–Alone" Methods for Impact Assessment

Preferred Methods

- Randomized experiments
- Quasi–experiments
- Multilevel regression methods

Alternative Methods

- Decomposition
  (proximate determinants model)
- Prevalence method

Note that the methods listed in Figure IV–1 exclude a number of methods that have been presented and/or applied elsewhere in the literature (United Nations, 1979, 1982, 1985; Chandrasekaran and Hermalin, 1985; Lloyd and Ross, 1989; Buckner et al., 1995). These are not covered in the present document for several reasons. Some of these methods (e.g., the standard couple–years of protection – SCYP, reproductive process analysis, and component projection methods) are based upon facility–level data, and are thus limited in their capacity to measure program impact at the population level.[11] The SCYP and reproductive process analysis methods all require data that are only occasionally available on a country–specific basis. Other methods (e.g., standardization, generic decomposition, and fertility projection/trend analysis) are relatively crude

[11] SCYP is not to be confused with conventional couple–years of protection (CYP), which is widely used to track program outputs.

methods that often lead to ambiguous conclusions regarding program impact in actual practice.[12] Finally, simulation is viewed as a generic method for analysis that may be meaningfully used in program and strategic planning and as a supplement to the more robust methods, but it is not especially useful as a stand–alone method of impact evaluation.

Accordingly, attention is limited in the present document to methods and approaches that are viewed by The EVALUATION Project as having the best prospects for producing relatively "clean" measures of program impact.

PREFERRED APPROACHES

Randomized Experiments

Description

It is widely accepted among evaluation researchers that the randomized or "true" experiment is the "gold standard" for measuring what has happened as a result of a program or intervention. The basic idea behind a randomized experiment is quite simple. In a randomized experiment, study subjects or groups are assigned to "treatment" and "control" groups randomly; that is, once the subjects or groups of subjects to be studied have been chosen, some are assigned to the treatment group and some to the control (or comparison) group at random. Given sufficiently large sample size, randomization enhances the chances of unambiguously isolating the effects of a program or intervention by distributing extraneous factors equally across comparison groups; that is, by ensuring that the treatment and control groups are equivalent with respect to all factors other than exposure to the program being evaluated. This conveys an enormous advantage over other methods of measuring program impact. It is for this reason that Rossi and Freeman (1993) refer to the randomized experiment as "the flagship" of evaluation.

Figure IV–2

Diagram of Two Commonly Used Randomized Experiments

Posttest–only control group design

Random Assignment:

	Time →	
Experimental Group	X	O_1
Control Group		O_2

Pretest–posttest control group design

Random Assignment:

	Time →		
Experimental Group	O_1	X	O_2
Control Group	O_3		O_4

Where:
X = program or intervention introduction
O = observations or measurements

Source: Campbell and Stanley, 1963

Design and Analysis

Two of the more commonly used randomized experimental designs are illustrated in Figure IV–2. In the first, the "posttest–only control group design," it is assumed that randomization has produced treatment and control groups that are equivalent, and it is thus necessary only to compare outcome measures for the treatment and control groups after the program has been operating for a sufficiently long period of time in order to assess the impact of the program or intervention being evaluated. In the second design, the "pretest–posttest control group design," measurements are taken for both treatment and control groups prior to program implementation and again after a period of time thought to be sufficient for the program to have had its intended impact. By taking "before and after" measurements, the researcher can subsequently correct for the fact

[12] It should be noted, however, that some of the excluded methods may be meaningfully used in conjunction with the more robust methods. Standardization, for example, is often used as a first step in an impact evaluation in order to determine the share of fertility change that is attributable to changes in demographic structure, since this share of fertility change clearly cannot be attributed to family planning program interventions. Similarly, simulation techniques may be used to supplement the information obtained from certain methods (see, for example, the discussion of multilevel regression methods in Section IV) by indicating the magnitude of change in outcome variables that may be expected from specified changes in program inputs or outputs.

that randomization may not have produced entirely equivalent groups.[13]

In both designs, the outcome measure(s) for the control group provide(s) an estimate of what would have been observed for the treatment group had the program under study not been implemented. In the posttest–only design, an estimate of program impact is provided by the difference in outcomes between the treatment group and the control group, plus/minus an error component that is taken into account as part of the statistical analysis; that is:

Impact = $(O_1 - O_2)$ +/– error

where: O_1 = outcome measure for the treatment group;

O_2 = outcome measure for the control group; and

error = design and measurement errors.[14]

In the pretest–posttest design, program impact is measured by the difference between the observed change in outcome measures for the treatment group less that for the control group, plus/minus error:

Impact = $(O_2 - O_1) - (O_4 - O_3)$ +/– error

where: O_1 and O_2 = pretest and posttest outcome measures, respectively, for the treatment group;

O_3 and O_4 = pretest and posttest outcome measures for the “control” group; and

error = design and measure–ment errors.

The Taichung experimental study described in Figure IV–3 illustrates how the randomized experiment can be used in an applied setting.

Figure IV–3

Illustrative Example of the Use of a Randomized Experiment for Program Impact Assessment

The Taichung experiment was designed to assess the impact of an effort to increase contraceptive awareness and use in the city of Taichung, Taiwan during the early 1960s. Local areas, or “lins,” in the city were randomly assigned (after geographic stratification with regard to density zones) to one of four experimental groups: (1) Full package, husband and wife: households in this group received home visits by health workers, mailings of information, and neighborhood meetings; (2) Full package, wife only: same intervention as for the first group, excluding the home visit to the husband; (3) Mailings only; and (4) No intervention other than family planning posters that were distributed throughout the city (i.e., the control group). Lins were allocated to experimental groups as follows: (1) n=427, (2) n=427, (3) n=768, and (4) n=767. Pre–intervention levels of contraceptive use were assumed to be equal across the randomized groups, and thus the posttest–only control group design was used.

The posttest observations of contraceptive acceptance rates (i.e., rates per 100 married women aged 20–39 years) for the four experimental groups for selected time periods (all density sectors combined) were as follows:

	Contraceptive Acceptance Rates	
Experimental Group	13+ months	29+ months
1. Full package—husband & wife	17	25
2. Full package—wife only	17	26
3. Mailing only	8	16
4. Control group	8	18
Total	11	20

Interpretation: Twenty–nine months after implementation, an increase in the contraceptive acceptance rate of 7 per 100 married women may be attributed to the “full package – husband and wife” intervention (i.e, calculated as the acceptance rate for this experimental group, 25 per 100, minus that for the control group, 18 per 100). Including husbands in home visits had no effect on contraceptive acceptance rates as may be inferred from the similarity in contraceptive acceptance rates for “husband and wife” and “wife only” experimental groups, nor apparently did the mailing of information.

Source: Freedman and Takeshita, 1969.

[13] This is of particular concern in small studies. As the sample size for the sudy increases, the likelihood that randomization will produce equivalent experimental groups also increases.

[14] The error component consists of both design effects and stochastic, or random, errors. Design effects refer to the biases introduced by factors on which the experimental groups differ despite randomization (assumed to be of minor importance in a randomized experiment, given sufficient sample size). Stochastic errors are assumed to be equivalent across experimental groups.

Several additional points regarding randomized experiments warrant mention. First, it should be noted that while random assignment of individual study subjects to experimental groups is quite common in clinical trials and smaller–scale studies involving individual program components (e.g., the effects of improved counseling on contraceptive continuation), individual assignment is more difficult and generally infeasible in large, population–based studies. In such studies, randomization is usually carried out at the group level; for example, at the level of villages, municipalities, districts, etc. The random assignment of groups of study subjects to experimental groups is illustrated in Figure IV–3.[15]

Second, it is possible that several program activities or interventions may be evaluated simultaneously in a randomized experiment by including multiple treatment groups in the design, one for each type or variant of "treatment." This "factorial" design is also illustrated in the example in Figure IV–3.

Third, it is not necessary for all treatment or benefits to be withheld from the control group in order for randomized experiments to be used. This is an important point, since in population–based studies of family planning program impact, it is difficult indeed to find a population without access to some family planning services (that is, a "pure" control group). What will be measured in such cases, however, is differential or incremental program impact; that is, the difference between the identifiable activity(ies) that constitute the intervention or "program" and other programs that may be operating simultaneously. Although this may seem an undesirable scenario to some, the fact that a large percentage of national populations in developing countries have access to some family planning services means that the impact of new programs will be incremental to those programs or services already in existence. In this sense incremental impact is an appropriate measure of what a program has accomplished.

Finally, it should be noted that randomized experiments are generic research designs as opposed to methods developed specifically to measure a particular type of outcome of family planning programs (e.g., methods designed specifically to measure fertility impact). Accordingly, they may be applied to assess program results at several different levels; for example,

- at the level of population–based outcomes (e.g., in terms of contraceptive prevalence or current fertility);
- at the level of program outputs (e.g., improvements in service accessibility or quality, an increase in numbers of new acceptors); and
- at the level of functional areas of service delivery (e.g., the effects of new staff training programs, supervisory systems, or clinic operational procedures on service delivery).

Program results may also be measured in relation to costs; for example, at the program output level, the increase in number of new acceptors may be related to the increase in costs for the different interventions or different packages. For the functional areas, the effects of two training programs may be related to their costs.

Strengths

The primary strengths of randomized experiments may be summarized as follows:

- Versatility

  Randomized experiments may be used to assess the results of program activities at several different levels in addition to overall impact.
- High internal validity

  This design is superior to all other designs for measuring program results in terms of minimizing threats to internal validity.
- Few assumptions required

  The primary assumptions required are that: (a) randomization has produced treatment and control groups that are equivalent, (b) influences external to the study affect both groups equally, (c) all treatment groups (or group members) receive the same "intensity" treatment, and (d) assignment to experimental group does not in itself alter the behavior of service providers or study subjects with respect to the outcomes under study.

[15] Note that where randomization is to be carried out at the group level, the ideal configuration is to have as many small–sized groups as possible.

- Relatively simple analysis

  Provided that the assumptions outlined above are valid, only relatively simple statistical tests are required.[16]

Limitations and Practical Considerations

Despite their theoretical attractiveness, randomized experiments have seen limited application in the evaluation of family planning programs, especially in the measurement of population–based outcomes. A number of reasons have been cited for this in the research literature (these are systematically reviewed in Bauman et al., 1994), some of which are well–founded, while others are arguable.

- Political or ethical sensitivities

  One factor often cited for not using randomized experiments is the political or ethical sensitivity related to withholding a desirable program from some segments of a population. While this may pose a dilemma in the short run, in many cases it may be neither possible nor prudent to implement a new program or intervention on a full–coverage basis. Testing programs on a limited basis and phased implementation of programs are quite common practices that lend themselves to the conduct of randomized experiments.[17] Thus, it is not necessarily the case that a randomized experiment will result in benefits being withheld to a greater degree than if the experimental study been not been undertaken.

- Time and cost

  Another argument against the use of randomized experiments is the time and cost involved. With regard to time, it should be recognized that program efforts may take time to mature, and thus if the primary evaluation objective is to measure medium– or long–term population–based outcomes, there is simply no alternative but to allow sufficient time for population–based changes to occur. In this regard, randomized experiments take no more time than other approaches to measuring program results.

  In terms of cost, whereas randomized experiments are thought to be expensive, the cost of a randomized experiment is not necessarily higher than a large–scale population–based survey such as the DHS (depending, of course, on the magnitude of the experimental study and, in particular, the number of posttest observations made).

- Generalizability

  Unless national level experiments are undertaken, the generalizability of findings of experimental studies in selected areas or subpopulations is often uncertain.

- Threats to validity

  Although less vulnerable than other study designs, randomized experiments are nevertheless subject to several potentially serious threats to validity. Among the more important of these are:

  - ➢ Contamination

    Contamination occurs when some or all of the control group is exposed to the intervention under study through such means as communication or migration between groups. This threat is illustrated in the Taichung experiment, where the increase in contraceptive acceptance rates in control lines is likely the result of diffusion of information about family planning to residents of such areas.

  - ➢ Confounding external influences

    One of the key assumptions of a randomized experiment is that treatment and control groups are exposed to the same external influences over the life of the study. In developing country settings, however, the presence of multiple international donors may well result in the introduction of new programs or initiatives during any given interval of time; for example, a 5–year program cycle. In such situations, it is possible that a control group for one intervention may be seen as being under–served and thus a logical target for another intervention. It is also possible that the treatment groups for one intervention may be used as a treatment group for another intervention, thus confounding attempts to measure the independent impact of the two interventions. External

[16] However, because of limited sample sizes and frequent violation of assumption, randomized experiments are often treated as quasi–experiments for analytic purposes, and the advantage of analytic simplicity is thus at least partially lost.

[17] In one sense, the inabliity to reach the entire target population for a program all at once provides an added rationale for the conduct of randomized experiments.

influences that affect the experimental groups differently undermine the validity of randomized experiments. Thus, in order for randomized experiments to be a realistic option in measuring program impact at the population level, it is essential to maintain control over the introduction of new interventions over the life of the experimental study.

➤ Variations in treatment

Especially in large–scale programs such as national family planning programs, programs are often implemented differently across geographic areas and/or service providers. For example, one or more program elements may be modified to meet local conditions, or the prescribed program simply may be implemented with varying levels of intensity in different areas. Thus, the measure of program impact will reflect the average impact of all program modifications and variations in intensity, instead of the program as it was designed. Where process evaluations are not undertaken concurrently in order to measure and understand such variability in implementation, this unmeasured variability might lead to misleading inferences as to the magnitude of program impact.

Threats to validity aside, there is at least one other practical consideration that may limit the utility of randomized experiments for measuring program impact: it may not make sense from a programmatic point of view to locate programs or interventions randomly. In fact, programs are often targeted at geographic areas or population subgroups for two diametrically opposed reasons: they are thought to be under–served or especially receptive to the program. While it is possible to conduct randomized experiments within such special populations, the likely result would be to dilute the impact of the program over the short– to medium–term. In such a situation, the need to demonstrate impact may be in competition with the ability to generate it.

To conclude, then, we strongly endorse the use of randomized experiments, but recognize that practical realities often limit their use, particularly in national–level impact studies. A case may be made, however, for the proposition that randomized experiments have been under–utilized in operations research and in studies involving program–level results, and should be the method of choice in such undertakings. Given the limitations of randomized experiments for measuring program results at the national level, however, alternative approaches need to be considered.

Quasi–Experiments

Description

The term "quasi–experiment" refers to a group of experimental research designs in which study subjects or groups of subjects are not randomly assigned. The most commonly–used quasi–experimental designs, "constructed control designs," follow the same logic and involve the comparison of treatment and control subjects or groups of subjects as in randomized experiments. In other designs, referred to as "reflexive control designs," treatment group subjects or groups of subjects serve as their own controls and time–series methods are used to measure net program impact (Rossi and Freeman, 1993). Though more vulnerable to threats to validity than randomized experiments, quasi–experiments do not require random assignment to experimental groups and therefore are generally more feasible than randomized experiments.

The numerous quasi–experimental research designs are discussed at length elsewhere (Campbell and Stanley, 1963; Cook and Campbell, 1979; Rossi and Freeman, 1993; Fisher et al., 1991). In this document, we focus attention on a design that has the widest applicability in assessing the impact of family planning programs: the pretest–posttest, non–equivalent control group design.[18]

[18] A number of potentially powerful quasi–experimental designs have been excluded from the discussion because they are unlikely to be widely applicable in the evaluation of family planning programs. For example, the time–series design is relatively strong where a reasonably long time–series data exist in few countries, and are largely limited to countries where population–based surveillance systems have been implemented (e.g., Matlab, Bangladesh and Cebu, Philippines). A second design, the regression discontinuity design, is perhaps the most powerful of the quasi–experimental designs available, but because the level of program "screening" required for the meaningful application of the design (e.g., the use of income or other eligibility criteria to choose program participants) is unlikely in family planning programs (which are based upon the notion of consumer choice), it is difficult to envision the circumstances under which this design would be applicable. Further details on these designs are provided in Rossi and Freeman (1993) and Cook and Campbell (1979).

Design and Analysis

The basic layout for the pretest–posttest, non–equivalent control group design is identical to that in the pretest–posttest randomized experimental design displayed in Figure IV–2, except that randomization is not used to assign study subjects or groups of subjects to experimental groups. Instead, one or more control (or comparison) groups are identified that are as similar as possible to the treatment group on as many factors as possible. In many applications, treatment and comparison groups are matched with respect to characteristics thought to be associated with the outcome under study (other than, of course, the program or intervention being evaluated). For example, subjects or population subgroups that are as similar as possible to the treatment group with respect to economic status, geographic location, ethnicity, and other characteristics might be purposively chosen to serve as comparison groups. Alternatively, geographic areas and/or population subgroups that are similar to the treatment area/population may be identified and a random sample chosen to serve as a comparison group.

As in randomized experiments, program impact is measured in the non–equivalent control group design by the difference between the change in outcome measures for the treatment group and that for the comparison group, plus or minus random error; that is,

Impact = $(O_2–O_1) – (O_4–O_3)$ +/– error

where: O_1 and O_2 = pre– and posttest measures, respectively, for the treatment group;

O_3 and O_4 = pre– and posttest measures for the comparison group; and

error = design effects and random measurement error.

In a quasi–experiment, it is of crucial importance to compensate for differences between treatment and control groups through the application of multivariate statistical methods. Even in matched studies, it is usually necessary to introduce statistical controls in order to control for differences in factors on which it was not possible to match. Experimental group differences that are not adequately controlled will be reflected in the design effect error component above and will directly influence the magnitude of the estimated impact. This is the primary disadvantage of quasi–experiments in comparison with randomized experiments: in randomized experiments, design effects are minimized by random assignment to experimental groups. The validity of quasi–experimental studies thus rests upon the effectiveness with which design effects can be minimized through matching and multivariate analysis.

Illustrative applications of the non–equivalent control group design are provided in Figures IV–4 and IV–5. Figure IV–4 displays results from a relatively strong variant of the design under consideration that features multiple observations of outcome measures both before and after program implementation. Figure IV–5 illustrates the more typical situation where single "pre–" and "post–" intervention measurements are made.

Strengths

The primary strengths of the non–equivalent control group design are:

- It provides an approximation to a randomized experiment when randomization is not possible.
- It is versatile. Like randomized experiments, quasi–experiments may be used to measure results at either the population or program levels.
- When properly designed, controlled, and analyzed, quasi–experiments can provide evidence of program impact that is nearly as strong as randomized experiments and stronger than most non–experimental studies.

Limitations and Practical Considerations

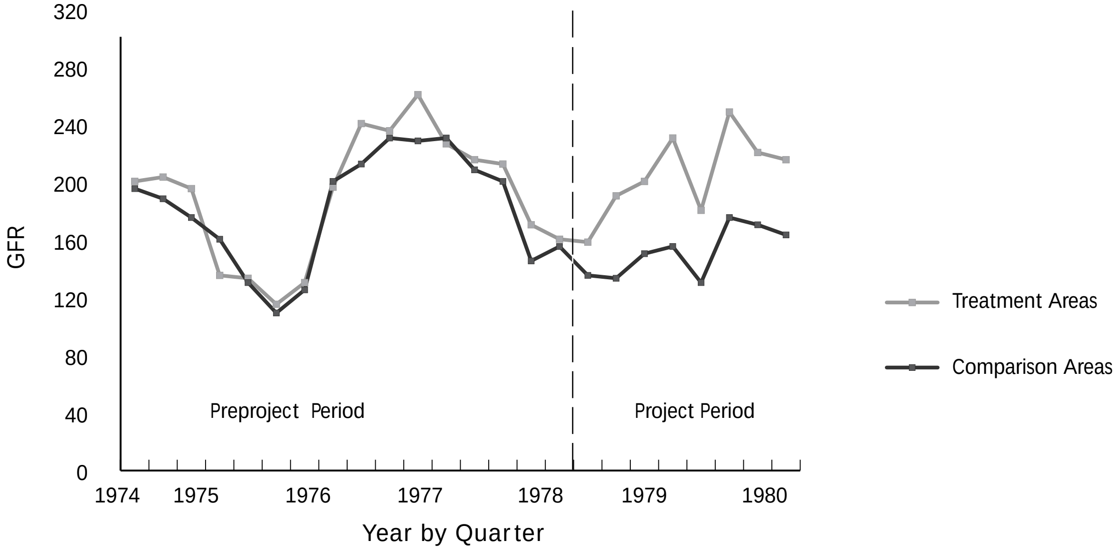
The non–equivalent control group design is subject to the same general assumptions and limitations as randomized experiments outlined earlier (other than those involving randomization).

In addition:

- The design is more vulnerable than randomized experiments to selection bias; that is, that differences in the characteristics of the experimental groups will be correlated with the outcomes under study, thus distorting the impact findings.
- It relies heavily upon multivariate statistical methods, and is thus sensitive to the use of appropriate statistical models and to the proper treatment of statistical estimation problems.

Figure IV–4

Illustrative Application of the Nonequivalent Control Groups Quasi–Experimental Design with Multiple Pretest and Posttest Observations, Matlab, Bangladesh, 1974–1980



Source: Phillips, J. et al., 1982, "The Demographic Impact of the Family Planning–Health Services Project in Matlab, Bangladesh," *Studies in Family Planning* 13 (5): 131–140.

As a practical matter, it is often possible in quasi–experimental studies to compensate for experimental group differences on key characteristics through matching and multivariate analysis. A lingering concern, however, is whether experimental groups differ on unobserved factors that influence the outcomes under study. Unlike the distorting effects of differences in factors that are observable/measurable and which can be accounted for through matching and the introduction of control variables in multivariate statistical models, factors that are unobservable (e.g., differential predisposition or motivation) cannot be compensated for in this fashion and can lead to misleading and/or biased estimates of program impact. This "unobserved heterogeneity" factor is in fact a concern in all study designs other than randomized experiments.

Considerable work has gone into the development of statistical methods for measuring and controlling the effects of unobserved factors, and these developments have been incorporated into the regression–based methods for measuring program impact that are discussed next in this chapter. While these developments have not been extensively used in conjunction with quasi–experimental research designs to date, there would not appear to be any reason why they could not be in future research.

Multilevel Regression Methods

Overview

Impact assessments based upon multilevel regression methods fall under the general heading of non–experimental or observational studies; that is, studies in which there are no experimental and control groups per se. Because the approach is non–experimental, the treatments vary from area to area as a result of decision–making processes that are beyond the control of the evaluation researcher. The criteria underlying program location or resource allocation decisions may or may not be known to the researcher.[19]

Multilevel regression methods are an extension of the multivariate areal regression methods that were rather widely used in the late–1970s through mid–1980s (e.g., Hermalin, 1975, 1979, and 1982; Poston and Chu, 1987; Chamratrithirong et al., 1986). The basic idea of a real regression is to try to demonstrate a statistical relationship between measures of family planning program activity/effort and selected outcome measures (e.g., contraceptive prevalence, total fertility rate) using geographic areas as the unit of analysis, while holding constant the effects of non–program factors such as age–sex composition, urbanization, female education and labor force participation, ethnicity, etc. The effects of non–program factors on the outcomes of interest are controlled through the application of regression methods. The approach attempts to answer the question "net of non–programmatic factors, do communities with a stronger family planning program presence tend to have higher contraceptive prevalence rates and/or lower fertility rates than those with lesser program presence?"

The multilevel approach extends the basic ideas of areal regression to the case where variables measured at different levels are used; for example, variables measured at the level of individual survey respondents, sample clusters, districts, and/or higher levels of aggregation. The use of data measured at different levels permits deeper insights into the implementing pathways through which family planning program inputs influence individual behaviors and the ways in which programs interact with other variables (e.g., the provision of services in villages with higher mean education levels having a greater effect on contraceptive use than in villages with lower levels of education).

Figure IV–5

Illustrative Application of the Nonequivalent Control Groups Quasi–Experimental Design with Single Pretest and Posttest Observations, Guatemala, 1983–1984

In order to assess the impact of three communications strategies designed to increase awareness and acceptability of vasectomy in Guatemala, "before and after" measurements were taken in four communities of similar socio–demographic characteristics. The three communications strategies were: (1) radio, (2) male promoter, and (3) both radio and male promoter. One community was chosen in which to implement each of the strategies to be tested, while a fourth was chosen as a control community. Baseline and follow–up survey data were collected for n=400 men of reproductive age in each of the four communities in June 1983 and again in July 1984, along with service statistics indicating the (monthly) number of operations performed in each community.

Communications program effects on knowledge and attitudes were assessed by comparing pre– and post–intervention measures of selected indicators. Logistics regression procedures were used to control for initial differences among the four communities and for possible "history" effects. Impact was assessed by comparing pre– and post–intervention vasectomy rates in the respective communities. Some of the key findings were as follows:

	Experimental Group							
	Radio & Promoter		Radio Only		Promoter Only		Control	
Indicator	Pre	Post	Pre	Post	Pre	Post	Pre	Post
% men interested in having a vasectomy	16.5	22.4	31.9	31.7	33.1	37.5	22.8	27.4
% interested men who had a vasectomy	0.5	0.8	1.5	2.0	1.0	3.0	0.3	1.0

Interpretation: The program did not increase interest in having a vasectomy in the target group (men who had heard of the operation, wanted no more children, were not already sterilized, nor were their wives); that is, the increases in the three communities did not differ significantly from the increase in the control community. In terms of actual vasectomy prevalence, only in the "promoter only" community did the change in prevalence exceed that observed in the control community by a statistically significant margin.

Source: Bertrand et al., 1987

[19] While most programs have guidelines or criteria for program resource allocation decisions, variations in program effort across geographic areas are also influenced by past program policies, the presence of multiple "players" in the family planning service delivery arena, uneven implementation of programs, and political factors. Thus, the relationship between current policies and program resources across geographic units may or may not accurately reflect current resource allocation priorities.

Because of these desirable features, as well as the wide availability of individual–level data from surveys such as the World Fertility Survey and the Demographic and Health Survey, the multilevel approach has largely replaced the areal regression approach in the toolbox of evaluation researchers in recent years. Areal regression methods continue to be used, however, in cross–national analyses (see Bongaarts, 1993, for an illustrative recent application).

In this section, we describe three types of multilevel regression models that may be used for measuring program impact. The basic question that each model attempts to answer is summarized in Figure IV–6.

The Basic Cross–Sectional Model

The cross–sectional multilevel model seeks to assess whether there is a statistical relationship between family planning program variables (e.g., presence of a family planning clinic in the community, number of contraceptive methods offered within 30 kilometers of the community) and family planning outcomes, controlling for socioeconomic and other non–program factors. The model uses randomly chosen communities or other areal units (e.g., municipalities, districts, provinces, etc.) and samples of individual women/couples from each community as units of analysis.

Figure IV–6

Primary Questions Addressed by the Principal Regression Modeling Approaches for Measuring Program Impact

Model	Question
Basic cross–sectional	Is there a statistical relationship between program and outcome variables (e.g., contraceptive prevalence, fertility) when the effects of other observed factors influencing the outcome(s) are con–trolled statistically?
Random effects model	Is there a statistical relationship between program and outcome variables when the effects of other observed factors influencing the outcome(s) and unobserved factors that jointly influence program and outcome variables are controlled statistically?
Panel model	To what extent are observed changes in outcome variables associated with changes in program variables when the effects of initial levels of and changes in other factors (both observed and unobserved) are taken into account?

To illustrate, we consider a relatively simple cross–sectional model relating program inputs to fertility behavior. A typical model might include three types of variables measured at two levels (i.e., at the individual/household and community levels):[20]

- factors specific to individual women and households (e.g., age, parity, education, demand for children, household assets, family structure, etc.);
- factors that are specific to communities or other population aggregations, but are common to all households and individuals within the community (e.g., environmental conditions, community infrastructure, labor market conditions, etc.);[21] and
- community–level measures of family planning program strength, which are also assumed to be common to all households and individuals in the community (e.g., presence of a fixed clinic providing family planning services in the community, number of family planning methods available at outlets within a specified distance of the community, quality of services).

[20] Individual– and household–level variables are usually treated as being at the same level. Models with more than two levels are possible, but the computer software currently available is limited in certain ways (e.g., the MLn software package uses approximations in 3– and 4–level models and can handle only dichotomous and continuous outcome variables). The interested reader is referred to Woodhouse (1995) for details on the MLn package.

[21] A fourth type of variable is often used in multilevel models: community–level variables that are derived by adding or averaging over individual observations within sample communities. Examples are mean household income or the proportion of households with electricity. This type of variable serves the same purpose as the community–level variables described in the text and has been omitted from the discussion in order to simplify the presentation.

The basic model may be written as:

Outcome Variable = Program Factors + Individual Factors + Community Factors + Interactions Among Factors[22] +/– Error[23].

Of primary interest for program evaluation purposes are the magnitude and statistical significance of the regression parameter(s) of the program variable(s) included in the model and the interactions between program variables and other variables. The regression parameters for the program variables indicate the strength of association between program measures and individual–level outcomes (e.g., contraceptive use, fertility) when the effects of other factors included in the model have been controlled; the interaction terms provide information on whether the program had a larger impact on certain population subgroups than others. For example, the inclusion of interaction terms in the model might allow the evaluator to test whether the provision of CBD points had a greater influence on the contraceptive behavior of lower–income women (who may have less access to fixed facilities) than higher–income women.

An illustrative application of the cross–sectional model to the measurement of the effects of selected characteristics of the family planning supply environment on contraceptive use is provided in Figure IV–7.

An important limitation of the cross–sectional model is the difficulty entailed in sorting out cause–effect relationships from measurements taken at one point in time. In the cross–sectional model, an observed positive relationship between family planning program variables and outcome variables (e.g., contraceptive use, fertility) may be interpreted as meaning that the program has influenced or caused the observed outcome, net of the other variables included in the model. However, it

Figure IV–7

Illustrative Application of the Cross–Sectional Multilevel Model, Thailand

Study Design Entwisle et al. (1984) used data from the second round of the Thailand Contraceptive Prevalence Study (CPS2) and multilevel regression methods to assess the effects of availability of family planning program outlets on the likelihood of contraceptive use in rural Thailand. Data collected from 4,956 rural women aged 15–44 years who were married or in union at the time of the survey were used in the analysis. Sample villages were classified into three groups on the basis of their proximity to different types of facilities providing family planning services: (1) villages located near (i.e., within 4 km.) a district health center, (2) villages located near a tambol (i.e., municipality) health center, and (3) villages located near neither type of facility. Individual–level variables used in the analysis included age, education, and desire for more children.

Results The regression results indicated strong effects of both family planning service availability and individual–level variables, as well as strong interactions between service availability and desire for more children and education. The adjusted odds (calculated from the regression results) of using a modern contraceptive method among women who desire no more children (in comparison with those desiring more children) are shown below for categories of the age and service availability variables.

	Age		
Service Availability	15–24	25–34	35–44
Near District Health Center	1.90	4.45	5.29
Near Tambol Health Center	2.07	4.06	7.40
> 4 km. From Either	2.04	1.72	4.77

Interpretation The strongest effects of service availability on contraceptive use were observed among women aged 25–34 years, with women residing near district or tambol health centers more than twice as likely as women more than 4 km. from either to have been using a modern contraceptive method at the time of the survey when other factors are controlled statistically.

Source: Entwisle et al., 1984

[22] The term "interaction" refers to the statistical dependence of a given factor or variable on other variables; for example, the effects of adding CBD points might depend upon the degree of access to fixed clinics in a given setting. Interactions between program and individual factors, between program and community factors, and between individual and community factors are possible.

[23] In multilevel models, both community– and individual–level errors are present. For the more statistically inclined reader, the cross–sectional multilevel model is written out in the form of a regression equation in Appendix A. The reader will note the specification of both community– and individual–level error terms in the regression equations.

is also possible that the causal pathway actually runs in the opposite direction; that is, that demand for health and family planning services caused services to be located in areas where contraceptive use was predisposed to be high.

Program managers often justifiably locate clinics or other types of outlets so as to meet existing demand for services and/or select locations with more advantaged populations or superior infrastructure (e.g., roads, electricity, etc.). Such factors have no doubt influenced location decisions in many family planning programs, particularly at the early stages of program development. To the extent that the populations served by these facilities are predisposed to higher contraceptive use and lower fertility, there is a danger that impact evaluations based upon cross–sectional data may overstate the actual level of program impact.

Note that the way in which programs are "located" may also lead to underestimating program impact from cross–sectional data; for example, where programs are targeted at areas of high fertility and mortality and/or low contraceptive prevalence.

The central issue is whether family planning program variables should be viewed as "endogenous" or "exogenous" variables in the regression equations.[24] If programs are implemented uniformly or randomly across sub–national units, then program variables may be viewed as being exogenous factors or variables without risk of estimation bias. However, if programs are implemented on a non–uniform or non–random basis according to some type of decision–making process, then the endogeneity problem comes into play. If so, this would result in a violation of the standard regression assumption that variables in the equation and error terms are uncorrelated. Of particular concern is whether "unobserved" factors that are correlated with the outcomes of interest may have influenced program location decisions. The crucial point is this — if program "location" decisions are made on the basis of factors that cannot be measured and controlled in a statistical model, one may obtain inconsistent and/or biased estimates of program impact. The interested reader is referred to Bollen et al. (1992 and 1995) for fuller discussions of these issues in the context of family planning program impact evaluation.

In assessing impact using multilevel regression models, it is essential to address the issue of endogeneity. Two approaches for doing so, multi–equation random effects models and fixed effects panel models, are described below.[25,26]

The Multi–Equation Random Effects Model

A number of approaches have been developed for dealing with the statistical problems resulting from unobserved variables in cross–sectional data. The interested reader is referred to Bollen et al. (1992, 1995) and Mroz and Guilkey (1992) for detailed discussions and appraisals of alternative approaches. Here, we focus attention on one approach that The EVALUATION Project finds particularly promising: an adaptation of the random effects model to the types of data that are typically available in DHS–type surveys supplemented by selected family planning program information.

Although random effects models are generally associated with longitudinal (as opposed to cross–sectional) data, recent methodological

[24] Endogenous factors or variables are independent or predictor variables that are determined by the same set of factors or the same decision–making processes that determine the outcome variable being studied. Exogenous factors, by contrast, are variables that are not determined by factors that also influence program variables. For example, labor force participation is endogenous in a regression equation predicting contraceptive use, since both labor force participation and contraceptive use are "choice" variables for individual women that are influenced by common factors (e.g., education, household size and structure). The community wage rate for women would be an example of an exogenous factor in such a regression, since this variable is determined by factors that are not subject to individual choice.

[25] It will be noted that the basic cross–sectional model described above can also be estimated within a random effects framework depending upon the assumptions made about the error terms.The distinguishing feature of the random effects approach described in the next section is the use of multiple equations that are estimated simultaneously.

[26] Note that other approaches are also available but are viewed by The EVALUATION Project as being less promising than the approaches presented in this publication. The interested reader is referred to Mroz and Guilkey (1992) and Bollen et al. (1995) for critical appraisals of these.

developments have extended the basic ideas for use with cross–sectional data sets where certain types of information:

- have been collected retrospectively in the survey questionnaires (e.g., retrospective birth histories in DHS);
- can be obtained by "backdating" from information gathered in DHS questionnaires and Service Availability Modules; or
- are available from other sources and can be used to supplement the information gathered in DHS–type surveys (Mroz and Guilkey, 1992; Newman, 1988).

The random effects model addresses the endogeneity problem by attempting to control statistically the effects of unobserved factors that jointly influence program location decisions and the outcomes under study. The basic idea is to estimate the distribution of such factors and control for them statistically in the regression model. The statistical procedures for accomplishing this are, however, quite involved and beyond the scope of this document. The interested reader is referred to Bollen et al. (1995), Mroz and Guilkey (1992), Guilkey and Cochrane (1994), and Newman (1988) for details.

The model also uses a structural equation approach in which the influences of "observable" community–level factors on the location of program resources in sample clusters are measured in one set of regression equations and the effects of program variables on the outcome(s) of interest (e.g., fertility, contraceptive use, etc.) in another equation. Both/all equations are estimated simultaneously using a full–information maximum likelihood (FIML) estimation procedure. The objective is to remove the influences of unobserved factors that may have influenced program location decisions from the equation relating program variables to outcomes.[27] Of primary interest for program evaluation purposes is the magnitude and statistical significance of the coefficients for the program variables in the outcome equation.

The model may be written as:

Program location equation(s):

Program Variables =
Observed Community Factors +Unobserved Community Factors +/- Error.

Outcome equation:

Outcome Variables =
Program Variables + Individual Factors + Unobserved[28] Community Factors +/- Error[29].

Two illustrative applications of the modified random effects model are presented in Figures IV–8 and IV–9. In the first example, the model is used to assess the impact of selected characteristics of the family planning supply environment on a "current–status" outcome (current use of modern contraceptive method) in Zimbabwe (Guilkey and Cochrane, 1994). In the second example, a somewhat more elaborate model is estimated for rural Tanzania. It takes advantage of retrospective birth history data (available in standard DHS surveys) and length of time that contraceptive methods have been available in sample clusters visited in connection with the DHS Service Availability Module to estimate program impact on fertility during the 20–year period prior to the 1991 Tanzania DHS (Angeles et al., 1995).

Note that in both examples the results are presented in the form of "policy simulations." The simulations provide a means of expressing multivariate results in a form that is readily understandable by program managers and policy makers; that is, in the form of answers to "what if . . . ?" questions. For example, what would the total fertility rate have been during a given period if all sample clusters had had the full range of family planning service and methods available? The simulation results are derived from the regression results by substituting values for selected program variables into the regression equations to depict different programmatic conditions (e.g., full coverage of villages by CBD points, adding a trained family planning service provider to

[27] The basic idea here is to try to include as many of the factors that may have influenced program location/ allocation decisions as possible in the first equation in order to control for their effects.

[28] Note that a term for individual–level unobservables may also be added to the equation and measured using the same methods used to estimate the effects of community–level unobservables. They are not considered here in order to simplify the presentation.

[29] The model is written out in regression format in Appendix A.

Figure IV–8

Illustrative Application of the Random Effects Model, Zimbabwe

Study Design

The modified random effects model was recently applied to data from the 1988/89 Zimbabwe Demographic and Health Survey and the 1989/90 Zimbabwe Service Availability Survey in order to measure the effects of access to family planning services on contraceptive use. Survey data on 2,050 currently married women were used in the analysis, along with community–level and family planning service data for 167 communities (i.e., DHS sample clusters). The service data contained measures of physical access to service delivery/supply points, as well as a series of measures of service delivery system preparedness and functioning (e.g., presence of electricity and running water, numbers of staff trained in family planning, availability of contraceptive methods and supplies, and courteousness of staff). The statistical model employed consisted of a system of four equations. The outcome variables for the four equations were: (1) probability that a survey respondent has had "r" births over the course of her reproductive career, (2) probability that a survey respondent has had "r" infant/child deaths over the course of her reproductive career, (3) fertility intentions, and (4) current contraceptive method. The independent variables in the four equations consisted of exogenous individual– and household–level variables, selected community characteristics, and a set of variables measuring various aspects of the supply environment for family planning services (measured at the community level). Each equation also included a parameter representing unobserved factors hypothesized to influence all four outcome variables. The equations were estimated simultaneously using a full–information maximum likelihood estimator.

Results

The following individual–level variables showed statistically significant effects on the use of modern contraceptives in the reduced form equations: respondent's age, religion, years of education, and whether the respondent resided in a commune. Two community–level variables also emerged as significant determinants of contraceptive use: educational opportunities available in the community and the presence of a CBD point in the community. Most of the unobserved heterogeneity parameters were statistically significant, indicating that their omission from the model would have resulted in biased estimates of the effects of the other variables in the model. Simulations based upon the regression results revealed the following estimated effects of CBD points:

Proportion Using

Variable and Condition	No Method	Modern Method	Traditional Method
Actual Values	.47	.44	.09
CBD Point in Community	.44	.47	.09
No CBD	.50	.40	.10

Interpretation

The strongest predictors of modern contraceptive use were individual–level characteristics. Of the family planning program variables tested, only the presence of a CBD point in the community had a significant effect on modern contraceptive use. Simulations indicate that if CBD points were to be established in every community, modern–method contraceptive prevalence would be expected to increase by about 7 percent from observed levels (from .44 to .47) and by 17 percent from the level that would prevail if there were no CBD points (from .40 to .47) when the effects of other factors are taken into account.

Source: Guilkey and Cochrane, 1994.

Figure IV–9

Illustrative Application of the Random Effects Model to the Measurement of the Fertility Impact of Family Planning, Rural Tanzania, 1969–1991

Study Design

The cross–sectional random effects model was recently applied to data from the 1991 Tanzania DHS and accompanying Service Availability Module. The focus of the analysis was on estimating the fertility impact of family planning in Tanzania over the 1969–1991 period, and was based upon household survey data from 5,215 women residing in 242 rural sample clusters who were under age 35 in 1991. Information on the locations of health facilities, family planning service availability and time of initiation, and other service delivery characteristics were provided by the SAM. Because reliable data series were not available in Tanzania, it was necessary to estimate fertility and child mortality levels and trends from the retrospective histories gathered in the DHS and to establish the dates of initiation of family planning services by "backdating" from the DHS SAM data. Historical figures on government health expenditures were, however, available and were used in the analysis.

The outcome variable used in the study was the probability that a survey respondent had a birth in any given year *i* during the study period. The following variables were included as predictor or independent variables: (1) age of respondent in year *i*, (2) education level, (3) whether there was a hospital, health center, and dispensary located within 30 km. of the sample cluster in year *i*, (4) the district–level child mortality rate for year *i*, (5) whether family planning services were available at hospitals, health centers, and dispensaries located within 30 km. in year *i*, (6) the duration of family planning service availability at hospitals, health centers, and dispensaries in year *i*, and (7) whether family planning services were available at hospitals, health centers, and dispensaries located within 30 km. when the respondent was 12 years old.

Since dispensaries and (to a lesser extent) health centers had been deployed in areas with high child mortality over the past 10–15 years as a matter of policy, it was crucial to account for non–random program placement in the analysis. Accordingly, three separate equations were estimated in which the outcome variable was whether or not there was a hospital, health center, and dispensary (respectively) located within 30 km. of cluster in year i. Predictor variables in these equations included: (1) the district–level child mortality rate in year *i*, (2) whether family planning service were offered by other facilities located within 30 km. in year *i*, (3) government expenditures on health in year *i*, and (4) the population of the district in which the sample cluster was located in year *i*. These equations were then estimated simultaneously with the outcome (i.e., conception) equation in order to estimate the impact of the Tanzanian family planning program while controlling for other observed and unobserved factors. A full information maximum likelihood estimation procedure was used in the analysis.

Results

The estimated impact of the family planning program effort over the 1969–1991 period is summarized below in the form of policy simulations.

	Mean Annual Fertitlity Rate	Mean Children EverBorn	Pct. of Scenario 1
Actual (observed) values	0.181	4.16	.95
Scenario I: No family planning services	0.189	4.35	1.00
Scenario II: FP only in hospitals	0.158	3.63	0.83
Scenario III: FP only in health centers	0.165	3.78	0.87
Scenario IV: FP only available in dispensearies	0.166	3.82	0.88
Scenario V: FP available in all types of facilities	0.120	2.76	0.63

Interpretation

If family planning had been available continuously at all three types of facilities over the study period (with all other factors held constant at observed levels), annual birth probabilities and the mean number of children born per woman would have been 37 percent lower than those observed for this period (0.120 versus 0.181).

Source: Angeles et al., 1995.

each health center, etc.) and comparing the resulting "expected" values of the outcome variable(s) under study with the results based upon actual values for these variables measured in a survey.[30]

The Fixed Effects Panel Model

The second alternative approach to dealing with the problem of unobservable factors influencing program placement decisions, the fixed effects panel model, extends the basic ideas of the cross–sectional model to the case where observations are obtained for the same sample of individuals or communities at two or more points in time. Data sets where the same sample of individuals are followed over time are relatively rare. However, where communities or clusters are used as the units of analysis, much of the required data may be obtained from successive Demographic and Health Surveys that include Service Availability Modules undertaken in the same sample clusters.[31] This design addresses the question: do communities experiencing the greatest changes in the family planning supply environment between two points in time also show the greatest change in contraceptive use (or other outcomes), controlling for changes in other factors?

The basic model may be written as:

Changes in the Outcome Variable =
Changes in Program Factors + Changes in Individual Factors + Changes in Community Factors +/– Error.

A key advantage of the panel design is that it permits the use of a particular estimation procedure referred to in the literature as a "fixed effects" estimator. Under this approach, variables or factors are divided into two categories:

- those that vary during the time period for the evaluation study, or time–varying factors, and
- those that do not vary during the study period, or time–persistent or fixed factors.

Unobserved factors that may have influenced the allocation of program resources prior to the study period are treated as "fixed" in the model and "differenced out" of the regression equations,

Figure IV–10

Illustrative Impact Evaluation Design Using the Fixed Effects Panel Model, Tanzania

The multilevel panel design will be used to measure the impact of National Family Planning Program (NFPP) efforts during the 1991–96 period in Tanzania. The period covered by the impact evaluation coincides rather closely with the project period for the USAID/Tanzania Family Planning Services Support (FPSS) Project (1990–97). A Demographic and Health Survey was conducted in 1991/2 and another is planned for 1996. In part to compensate for the limited routine service statistics available, a smaller–scale interim or "mid–term" DHS (known as the Tanzania Knowledge, Attitudes, and Practices Survey, or TKAP) was undertaken in 1994. All three survey rounds (1991/2, 1994, and 1996) have been or will be conducted in the same sample clusters and will include Service Availability Modules to provide measures of program strength/activity (and changes therein) at the cluster level.

The focus of the analysis will be on assessing the nature and magnitude of changes in program service delivery at the community (or cluster) level during the 1991–96 period and the role that such changes play in influencing contraceptive behavior, fertility levels, and other outcomes of interest, controlling for the effects of changes in other factors.

The 1991/92 DHS estimated unmet need for family planning at 30 percent of currently married women. In recent years FPSS has invested heavily in training, as well as commodities and logistics management. Thus, it is anticipated that large effects will be observed in subsequent surveys for "supply environment" variables such as "presence of trained staff," "number of stock–outs in last six months," and "number of contraceptive methods offered at the facility." It is hypothesized that increases in contraceptive use and decreases in unmet need for family planning will be larger in areas where the greatest improvements in the family planning supply environment have taken place.

[30] Note that in the example in Figure IV–8 based upon Zimbabwe data, an equation for program placement is not included in the model, and it is thus necessary to assume that the family planning program is exogenous. This type of model is useful in examining the contributions of programs to contraceptive use that result indirectly from program effects on fertility intentions.

[31] For example, successive rounds of DHS have been undertaken in the same sample clusters in Morocco (1987,1992, and 1995) and Tanzania (1991 and 1994), albeit for half–samples in Morocco in 1995 and Tanzania in 1994. In Morocco, an attempt was made to re–interview the same women in 1995 as in 1992, resulting in a "true" panel of individual women.

thus reducing the risk of estimation bias.[32] This "differencing" procedure is illustrated in Appendix A, where the fixed effects panel model is written out in regression format.

The logic of the multilevel panel design and the types of data needed for its application are illustrated in the case of Tanzania in Figure IV–10.

Data Requirements

The data requirements for multilevel models are rather demanding. The following types of data are needed in order to construct sound cross–sectional models of family planning program outcomes:

- Household– and individual–level survey data (e.g., from DHS–type surveys) on:
  - demographic and economic characteristics;
  - fertility preferences and intentions; and
  - current contraceptive use, fertility, and/or other "outcome" measures.
- Information on community–level determinants of fertility and fertility demand (usually obtained from community–level surveys):
  - labor market conditions and wage rates;
  - community infrastructure; and
  - demographic indicators for prior years (see below for further discussion).
- Information on the supply environment for family planning services, usually obtained from facility surveys or censuses such as the DHS Service Availability Module, program statistics, and/or community–level surveys:[33]
  - number and types of health and facility planning facilities within a fixed distance of each sample cluster (e.g., 30 km.);
  - services and contraceptive methods available;
  - length of time services and methods have been available; and
  - measures of service quality.

Especially important in the random effects models is the availability of information on community–level characteristics that may have been important determinants of prior program location/resource allocation decisions (e.g., fertility and mortality levels, standard of living indicators, etc.).

Few (if any) surveys collect the full range of data required for meaningful applications of this approach. However, Demographic and Health Surveys that include a Service Availability Module (SAM) generally contain a good deal of the information needed. These can often be supplemented with the additional program and community–level economic data needed at relatively low marginal cost in order to provide an adequate database for estimating multilevel models. Other surveys with comparable information content may also be used.

The data requirements for applying the fixed effects panel model are largely the same as for the cross–sectional model, with the exception that data at two or more points in time on the same individuals or the same clusters are needed. Ideally, DHS data collection in a given country would be timed to correspond to cycles of family planning program activities (e.g., surveys at the beginning and end of a 5–7 year cycle). In this way, multipurpose surveys such as the DHS could serve as the basic mechanism for gathering the required information for measuring impact during a given program cycle. However, it is not necessary that the timing of rounds of DHS correspond exactly to program cycles. As survey rounds accumulate, the time frame covered by the multilevel panel study could be extended to compare program effects across different stages of program implementation. In addition, conducting cross–sectional multilevel analyses at different time points may be used to show that different inputs are important to achieving impact at different

[32] The model is focused on measuring the population repsonse to changes in the program during the period of time considered in the evaluation study. Program resource allocation decisions and investments made at earlier points in time are assumed not to be causes of changes in outcome measures during the evaluation study period except through delayed or lagged responses.

[33] The Population Council's Situation Analysis is another example of the type of facility survey that may be used to gather much of the required program–level information.

stages of program development (effects which may not show up when aggregating across program phases).

Strengths

The multilevel regression approach has a number of strengths:

- Since the approach relates program input measures to outcomes at the community level, it permits the measurement of impact of the program as it is actually implemented.
- It does not require an experimental design.
- It provides more detailed information on the pathways through which programs influence contraceptive behavior than any other approach.

Limitations and Practical Considerations

There are also several important limitations and constraints in the use of multilevel regression models:

- The approach is demanding in terms of data. Practically speaking, large–scale population–based surveys comparable to the DHS are required. This is especially true when panel models are to be estimated, as the unit of analysis for such analyses are communities or clusters. Unless a large enough sample of clusters is available, the study design may lack sufficient statistical power to detect program effects of the magnitude that are likely to occur over relatively short periods of time such as five years. For the cross–sectional model, the availability of community–level data on factors that may have influenced prior program location decisions is crucial if unbiased estimates of program impact are to be obtained. Measures of program activity, which usually are derived from facility–based surveys, are also required.
- The method is sensitive to the use of appropriate statistical models and to the proper treatment of statistical estimation problems. Suitably trained personnel and appropriate computer software are also needed in order to carry out multilevel analyses.
- Developing and estimating the statistical models require knowledge of the setting. Because the specific variables included in the models will vary from case to case, there is no standard set of variables to apply. Knowledge of the social setting and evolution of health and family planning services is needed to inform the selection of appropriate community level and program variables.
- The models are sensitive to the timing of program investments in relation to the period of observation in impact assessments. Since some specifications of the random effects model relate family planning program indicators to outcome measures in the recent past (see Figure IV–9 for an illustration of this point), accurate information on the timing of past program changes is crucial to the accurate estimation of program impact. Similarly, since the fixed effects model relates changes in program variables to changes in outcome indicators during specific time intervals, the absence (or minimal levels) of change in program measures during a particular period studied will logically result in estimates of minimal program impact.[34] The fixed effects model will also not pick up on program investments made prior to the period under study. For example, if a country had made extensive investments in family planning in the 1970s and early 1980s and had basically maintained strong support of these activities thereafter, the country might exhibit fairly high levels of contraceptive prevalence. Yet if one were to apply the fixed effects model with data from the late 1980s and early 1990s, the analysis might show relatively little impact, because the strong investment from earlier years would not be reflected in the measures of program change. This phenomenon might explain, in part, the relatively small program impact found in the Gertler and Molyneaux (1994) analysis of the Indonesian family planning program during the mid–1980s. Thus, caution needs to be exercised in applying the fixed effects model in countries with relatively advanced programs, where new

[34] In one sense, this is a generic evaluation design issue; an analogous problem arises in attempting to measure the outcome of a randomized experiment too soon after program inplementation. It is highlighted here in connection with the fixed effects model to emphasize its importance when interpreting the results of impact studies based upon this approach.

program investments during recent periods may be relatively modest.[35]

Despite these limitations, the multilevel approach constitutes a feasible and potentially powerful methodology for assessing program impact given the availability of DHS–type data on a recurrent basis in many countries, the growing proliferation of powerful microcomputers and appropriate computer software, and growing numbers of researchers with the methodological skills required to carry out multilevel modeling.

ALTERNATIVE APPROACHES

What options are available when it is not possible to implement any of the preferred approaches outlined in the preceding section? In this section,

Figure IV–11

Summary of Preferred Approaches to Measuring Family Planning Program Impact

Approach	Observations
Randomized Experiments	The “gold standard,” but generally impractical due to difficulties in establishing and maintaining controlled experimental conditions in national–level studies.
Quasi–experiments	More practical since random assignment is not required. However, more vulnerable to selection bias and other threats to validity than randomized experiments. Use of quasi–experimental designs along with methods to control for unobserved heterogeneity holds considerable promise.
Multilevel Regression Methods:	
Basic Cross–sectional Model	The least demanding of the preferred non–experimental approaches. Entails assessment of relationship between variables measuring family program presence/activities and outcomes based upon cross–sectional household and facility data gathered from the same geographic areas (e.g., DHS sample clusters). Key limitation —danger of biased estimates where programs are not implemented uniformly or randomly across geographic areas.
Multi–Equation Random Effects	Refinement of the basic cross–sectional model that provides a way to account for the possible endogeneity of program location/placement. Requires additional data on factors or variables that might be correlated with program placement decisions. Statistical estimation procedures are complex.
Fixed Effects Panel Model	Given household and facility data from the same geographic areas at two or more points in time, provides an alternative (and computationally simpler) method for addressing the problem of the endogeneity of program placement. Because sample clusters are the unit of analysis, generally has less statistical power than the other models.

[35] One solution to this problem would be to build “time–lags” into the fixed effects model. This is feasible, however, only where three or more rounds of household survey data with accompanying measurements of program variables are available.

two less preferred but nevertheless useful alternatives are presented:

- a specific type of decomposition and
- the prevalence method.

The decomposition approach requires the collection of data at two or more points in time, while the prevalence method is cross–sectional in nature.

Decomposition (Proximate Determinants Model)

Description

Strictly speaking, decomposition is a generic technique of demographic or epidemiologic analysis as opposed to a method designed specifically for program impact evaluation purposes. Here, we focus on a specific method of decomposition, the approach proposed by Bongaarts and Kirmeyer (1980), which has several features that make it a more valuable tool for program impact assessment than generic decomposition methods.

As applied to the measurement of family planning program impact, the objective of decomposition is to determine what share of an observed change in fertility rates (e.g., the total fertility rate) between two points of time is attributable to changes in different factors. The method under consideration, which will hereafter be referred to as proximate determinants decomposition method (as it is based upon the well–known model of the proximate determinants of fertility – see Bongaarts, 1978), decomposes changes in the TFR into changes in: (1) proportions married or in union, (2) average length of postpartum infecundability, (3) contraceptive prevalence, and (4) abortion rates.[36] Program impact is indicated by a reduction in fertility levels during a particular period studied, a large share of which is attributable to increases in "program" contraceptive prevalence.[37] Although not provided for in the original model, estimates of program contraception may be obtained by using survey data on source of supply or service statistics to estimate what share of total contraceptive use is attributable to program versus non–program sources.

Design and Analysis

The usual source of data for the application of the method is two consecutive population–based surveys in a given population.[38] Other than the availability of data at two or more points in time and information on source of contraceptive services/supply, there are no special design requirements for the use of the method.

The measurement of program impact proceeds from the decomposition of total fecundity into components as proposed by Bongaarts and Kirmeyer (1980) and Bongaarts and Potter (1983). The underlying model is a multiplicative model in which the factors considered are expressed as an index measuring their fertility inhibiting effects; that is, as measures of the extent to which each factor contributes to the difference between total fecundity and total fertility. The model is depicted graphically in Figure IV–12.

In the cross–section, the model may be written as:

$$TFR = TF * C_m * C_c * C_i * C_a$$

Where:

TFR = observed total fertility rate;

TF = total fecundity rate;

C_m = index of marriage;

C_c = index of contraception;

C_i = index of postpartum infecundability; and

C_a = index of abortion.

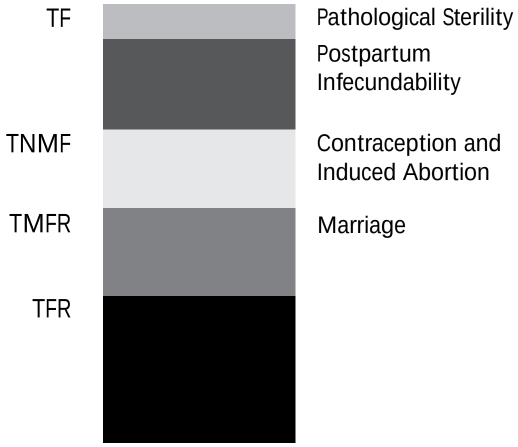
[36] The effects of variation in factors such as contraceptive use–effectiveness and sterility may also be incorporated into the mode, but setting–specific data on these variables are rarely available. Jolly and Gribble (1993) have proposed that an additional factor, fertility occurring outside of marriages or union, be included in the model in sub–Saharan African populations.

[37] Program contraceptive prevalence refers to contraceptive use that may be associated with program (as opposed to non–program) services and sources of supply. Non–program sources consist of the commercial and private sectors, and vary in relative importance from country to country.

[38] Population censuses also provide the required data on fertility levels, age composition and proportion married, but not on other proximate determinants of fertility.

Figure IV–12

Relationship Between the Fertility–Inhibiting Effects of the Proximate Determinants of Fertility and Selected Fertility Measures


TF
TNMF
TMFR
TFR
Pathological Sterility
Postpartum Infecundability
Contraception and Induced Abortion
Marriage

Source: Bongaarts and Kirmeyer. 1980. “The Proximate Determinants of Fertility.” in Foote, K., K. Hill, and L. Martin (eds.) *Demographic Change in Sub–Saharan Africa*. Washington, DC: National Academy Press.

When applied to successive rounds of survey data, changes in the indices for each of the proximate determinants are related to the observed change in the TFR in order to assess the contribution of each to the observed fertility change. For example, the proportional contribution of changes in contraception to an observed change in TFR is calculated as:

$$C = (\ln C_{c,2} - \ln C_{c,1}) / (\ln TFR_2 - \ln TFR_1)$$

Where:

ln = natural logarithm;

$C_{c,2}$ and $C_{c,1}$ = indices of contraception in survey rounds 2 and 1, respectively; and

TFR_2 and TFR_1 = observed total fertility rates in survey rounds 2 and 1, respectively.

As noted earlier, estimating program impact using this method requires the use of survey data or service statistics on sources of contraceptive supply or services to estimate the share of contraceptive use that is attributable to program sources. Illustrative results of the application of the method to successive surveys in the Philippines are presented in Figure IV–13.

Strengths

- The method requires only data typically available in DHS–type surveys.
- It is computationally simple.
- It is effective in controlling for effects of changes in proximate determinants of fertility.

Limitations and Practical Considerations

- Because of annual fluctuations in fertility levels, the method is sensitive to the particular years covered in the study; that is, quite different estimates might be obtained by varying the years covered.
- It is limited to the measurement of impact in terms of fertility.
- It does not provide direct measures of effects of program inputs; program inputs are inferred from changes in contraceptive prevalence (and estimated program contributions to changes in prevalence).
- In the absence of setting–specific data on factors such as contraceptive use–effectiveness, the method relies upon standard schedules, which may not accurately describe practices in the particular setting.
- One study (Reinis, 1992) suggests that the method may be invalid in settings where women use contraception primarily to stop childbearing once they have reached their desired family size (as opposed to space births), where delayed marriage is common, and where contraceptive use is most prevalent at older ages. This issue requires further investigation.
- The method is sensitive to: (1) accuracy of survey data on source of contraception and (2) definitions and reporting (in survey interviews) of program and non–program contraception.[39]

[39] These items have been among the least reliable items in Demographic and Health Surveys. In DHS–III surveys, the questions on this topic are being asked in a way that is expected to elicit more reliable information.

- The method provides only a measure of gross impact; that is, it does not account for source substitution and program catalytic effects (i.e., increases in non–program contraception that are the result of program promotional efforts).

Prevalence Method

Description

The final method considered, the prevalence method, provides a rough estimate of program impact when the collection of data at two or more points in time and the conduct of "posttest only" randomized experiments are not feasible. The prevalence method is a cross–sectional method designed to take advantage of the wide availability of survey data on contraceptive prevalence in developing country settings. Using survey data on contraceptive prevalence by source of supply or service (i.e., program vs. non–program), current age–specific fertility, and selected proximate determinants of fertility,[40] the method estimates the portion of the difference between potential fertility[41] and observed fertility that may be attributed to program contraception. This, in turn, may be converted into two estimates of program impact: (a) the reduction in fertility rates and (b) the number of births averted during a specified interval of time (usually a year) resulting from program contraception. The method is based upon the same model of the quantitative relationship between fertility and its proximate determinants described above in connection with the decomposition method (see Bongaarts, 1986, for a full presentation of the method).

Figure IV–13

Illustrative Application of the Proximate Determinants Decomposition Method to Successive Surveys in the Philippines

Survey Round and Reference Date of Estimates	TFR	C_m	C_c	C_i	TF
1978 Republic of the Philippines Fertility Survey (1973–1977)	5.60	0.599	0.778	0.761	15.77
1982 National Demographic Survey (1978–1982)	5.28	0.599	0.713	0.778	15.90

Change in TFR		Percentage point change in TFR contributed by changes in:			
Absolute	Pct.	Nuptiality	Contraception	Infecundability	Residual
–0.32	–5.6	0.0	–8.5	+2.1	+0.8

Interpretation During the period covered by these two surveys, total fertility declined 5.6 percent. The decline in TFR is explained entirely by an increase in contraceptive use. In fact, if only contraceptive use had changed during this period, TFR would have declined by 8.5 percentage points. However, changing levels of post–partum infecundability resulting from a decline in duration of breastfeeding exerted an upward influence on the TFR of 2.1 percentage points. Unspecified or residual factors also contributed to an increase of total fertility of 0.8 percentage points. The contribution of changes in abortion rates could not be assessed in this example due to the lack of appropriate data. As the survey data indicate that the public sector provided approximately 50 percent of contraceptive services and supplies during this time period, a reduction in the TFR of approximately 4 percent may be attributed to the public sector family planning program.

Source: Casterline et al., 1988.

Design and Analysis

The prevalence method is relatively non–demanding in terms of data requirements in the sense that the required data are normally gathered as part of DHS–type surveys. The basic data needed are:

- estimates of contraceptive prevalence for a specified point in time, by five–year age groups and source of supply;
- age–specific fertility rates for a given reference period;
- number of women of reproductive age, in five–year age groups; and
- total population size.

Setting–specific estimates of use–effectiveness, ideally by method and source, and age–specific

[40] Specifically, information on proportions of women of reproductive age married or in union and average length of post–partum insusceptibilty (often indexed by mean length of breastfeeding) are used. Information on abortion rates may also be used, where available.

[41] Potential fertility is defined as the level of fertility that would prevail in a given population in the absence of contraception.

fecundity rates are also useful, but only rarely available. Where such data are not available for a particular setting, standard schedules may be used (Bongaarts, 1986).

Under the method, program impact is measured by the share of the difference between potential fertility and actual fertility during a specified reference period (e.g., the 1 or 3 years prior to a survey) that is accounted for by program contraception. Two impact measures are normally produced in applications of the method: (1) the reduction in fertility rates (i.e., potential fertility minus observed fertility) attributable to program and non–program contraception, respectively, and (2) births averted by program and non–program contraception.

For illustrative purposes, Figure IV–14 provides estimates of effects on crude birth rates and of

Figure IV–14

Illustrative Results of the Application of the Prevalence Method, Selected Countries

				Gross Crude Rate Effect of:		Gross Births Averted by:	
Country or Area	Reference Year	Population Size POP (millions)	Observed Birth Rate CBR (per 1,000)	Programme Contraception GPCBR– CBR	Non–Programme Contraception NCBR–GPCBR	Programme Contraception BA (thousands)	Non–Programme Contraception BAN (thousands)
Africa							
Ghana	1978	11.37	46	1.7	0.0	19.5	0.0
Mauritius	1979	0.94	27	23.2	1.4	21.8	1.3
Tunisia	1977	6.01	35	6.5	1.1	39.2	6.9
Asia							
Hong Kong	1975	4.40	18	9.6	9.3	42.4	41.0
Indonesia	1979	148.09	33	9.6	2.0	1428.4	297.6
Malaysia	1979	13.67	31	11.2	3.7	153.1	50.6
Philippines	1979	47.68	32	6.9	9.1	329.4	432.3
Republic of Korea	1975	35.67	26	9.1	2.4	324.7	84.2
Singapore	1978	8.79	17	22.5	7.6	197.4	67.1
Thailand	1975	42.42	34	7.7	3.2	325.8	137.2
Latin America							
Donimican	1976	5.14	38	5.7	4.8	29.1	24.7
El Salvador	1976	4.26	42	3.3	7.1	14.0	30.1
Mexico	1978	64.09	36	9.6	10.6	616.5	681.4
Paraguay	1977	2.97	37	3.1	3.1	9.2	9.2
Oceania							
Fiji	1978	0.61	34	15.8	1.9	9.6	1.1

Interpretation for a selected country: In the absence of the national family planning program in Indonesia, the crude birth rate would have been 9.6 births per 1,000 population higher than the recorded 1979 CBR of 33. Non–program contraception reduced the CBR another 2.0/1,000. Over 1.4 million births are estimated to have been averted in 1979 due to program contraception, and nearly 300,000 more births were averted due to non–program contraception.

Source: Bongaarts, J. 1986. "The Prevalence Method." in United Nations *Manual IX Addendum*: *The Methodology of Measuring the Impact of Family Planning Programmes on Fertility*. New York: Department of International Economic and Social Affairs, pp. 9–14 (Table9).

births averted due to program and non–program contraception for selected countries for various years in the mid–1970s using the method.[42] Other illustrative applications are provided in United Nations (1986).

Note that although the method is cross–sectional by design, estimates from successive surveys provide a time–series of estimates that might be used to gauge trends in program performance over time. However, given data from successive surveys, the use of the proximate determinants decomposition method is likely to be more informative.

Strengths

The main strengths of the method are as follows:

- It does not require special studies to be undertaken. If it can be assumed that the standard schedules of use–effectiveness and fecundity are applicable to the population under study, the method requires only data that are normally available in DHS–type surveys.
- The computations are relatively simple.

Limitations and Practical Considerations

- If country–specific data on use–effectiveness (for both program and non–program contraception) and age–specific proportions of women who are fecund (not normally collected in DHS–type surveys) are not available, the method requires the assumption that standard schedules apply.
- It does not directly measure effects of program inputs; program inputs are inferred from changes in contraceptive prevalence (and estimated program contributions to changes in prevalence).
- The method is sensitive to: (1) accuracy of survey data on source of contraception and (2) definitions and reporting (in survey interviews) of program and non–program contraception.
- The method provides a measure of gross impact, but it does not account for source substitution and program catalytic effects (i.e., increases in non–program contraception that are the result of program promotional efforts).
- The method is limited to measuring impact in terms of fertility.

SUMMARY

In this chapter, three preferred and two next–best alternative approaches for measuring the impact of family planning programs were reviewed and critically appraised. The relative strengths and weaknesses of the approaches considered are summarized in Figure IV–15.

In overview, the recommended methods fall into two broad categories: experimental methods and survey–based methods.[43] The increasing reliance on survey–based methods for program evaluation is the result of two factors: (1) the fact that population–based outcomes are best measured from population–based data and (2) the relatively wide availability of data from large–scale, population–based surveys such as the DHS. The value of survey data for program evaluation purposes is enhanced when survey data collection efforts are strategically timed with respect to program cycles and are supplemented by program and community–level data for the same geographic units.

At the same time, experiments (both randomized and quasi–experiments) have a role to play in enhancing the chances of obtaining valid conclusions from program evaluation efforts, particularly in evaluations of the functional areas of family planning programs (i.e., operations research).

Ideally, different methods would be used to answer different questions as part of comprehensive program evaluations. However, the incorporation of different evaluation design components requires careful planning at the design stage of programs and new program cycles. Where there is a strong commitment to measuring impact, program administrators may choose to modify the manner in which activities are to be implemented to accommodate the meaningful measurement of program impact.

[42] Although relatively simple, the computations required in applying the method are rather lengthy. Accordingly, Figure IV–II illustrates the output that results from application of the method and its use in assessing the magnitude of program impact. The interested reader is referred to Bongaarts (1986) for full computational details.

[43] It should be noted that surveys are also often used as a means of data collections in experimental studies.

Figure IV–15

Summary of Characteristics of Methods for Measuring Program Impact

			Multilevel Regression			
Characteristic [a]	Random Experiments	Quasi Experiments	Single Survey	Panel	Decompo-sition	Prevalence Method
Exposure to threats to validity	Low	Moderate	Moderate	Moderate	Moderate	Moderate
Required assumptions	Few	Moderate	Moderate	Moderate	Many	Many
Ability to isolate program effects	Strong	Moderate	Moderate	Moderate	Moderate	Moderate
Volume of data required	Low	Moderate	High	High	Moderate	Low
Insights into causal pathways provided	Moderate	Moderate	Strong	Strong	Weak	Weak
Outcome indicators used	Flexible	Flexible	Flexible	Flexible	Fertility only	Fertility only
Degree of program control required	High	High	None	None	None	None
Feasibility where program activities are highly targeted	Low	Low	Moderate[b]	Moderate	High	High

[a] Methods are not compared in terms of cost in the table since costs will vary depending upon specific features of each application.

[b] Feasibility depends upon whether the factors underlying program targeting decisions can be incorporated into the regression equations as control variables.

Chapter V

Developing an Implementation Plan

- Defining the Institutions and Individuals Responsible for the Evaluation
- Establishing a Timetable for Specific Activities
- Budgeting the Evaluation

Chapter V

DEVELOPING AN IMPLEMENTATION PLAN

The previous chapters in this manual stress the technical aspects of designing an evaluation. However, an evaluation plan will be of little worth if there is not a clearly defined plan for its implementation. This chapter outlines the issues to be covered in developing an implementation plan for either program monitoring or impact assessment.

DEFINING THE INSTITUTIONS AND INDIVIDUALS RESPONSIBLE FOR EVALUATION

Generally, one institution takes the lead in designing and implementing the evaluation of a national family planning program, although it is often necessary to enlist the assistance of other service delivery or research organizations. In some cases the impetus for a large scale evaluation comes from the country itself; in others, it is a donor requirement. Perhaps the most common scenario is that program administrators and donors have a mutual interest in learning whether the program is on track and how it might be further strengthened.

The lead organization is often the major service provider in the country, especially if it is a governmental institution (e.g., the Ministry of Health). Alternatively, the private family planning association may take the lead, especially if it is a major player in service delivery and/or has a strong research/evaluation capability. Whoever has the prime responsibility for the evaluation, it is important to identify and involve other stakeholders.

Involving Key Stakeholders in the Planning Process

To maximize the benefit and utility of evaluating the national family planning program, it is important to include the major stakeholders in the process from the start. “Stakeholders” are all organizations or individuals who could potentially be interested in how the evaluation is carried out, what the results show, and how the information might subsequently be used. Also, the list should include any institutions expected to contribute data to the effort. Such potential stakeholders include the following:

- Official government offices responsible for monitoring population phenomena, especially in countries that have set demographic targets to be attained:
    - Ministry of Planning
    - National Population Council
    - Other
- Organizations that provide family planning services, including:
    - The Ministry of Public Health
    - The IPPF affiliate
    - Other major NGOs
    - Subsidized contraceptive social marketing programs
    - Private sector firms that market contraceptives
    - Associations of private providers (local OB–GYN society, midwives association)
- Donor agencies that support the program
- Women’s health and other advocacy groups

Ultimately, if the stakeholders do not perceive the data and analysis to be useful for the kinds of decisions they need to make about program design and implementation, the results may never be used for their intended purpose but rather may be ignored or discredited by those the evaluation is intended to assist.

Defining Technical Needs and Identifying Available Sources In–Country

A large scale evaluation of a family planning program, especially if it includes impact assessment, requires technical expertise in study design, preparation of data instruments, supervision of data collection, editing and processing of the information, data analysis, and report preparation.

A growing number of developing countries now have staff with expertise in these areas. Under ideal circumstances, appropriate technical staff will exist within the lead or collaborating organizations, and these individuals can be made available to work on the evaluation. However, it may occur that the appropriate individuals are already committed to other activities during the period of the evaluation, or the appropriate level of technical expertise is not available within the collaborating organizations. In this latter case, it is important to consider alternative sources of technical assistance in the areas outlined above (study design, data collection, analysis, etc.), such as:

- local research firms, especially those that specialize in social science research;
- local universities (e.g., departments of demography, schools of public health); and
- private consultants with research/evaluation background.

In countries where there is limited in–country technical expertise in social science research, it may be necessary to supplement the skills of local researchers with external technical assistance.

Establishing and Maintaining Effective Communication Channels

It is important for the lead organization to maintain an honest dialogue with the larger group of stakeholders, not just expect the others to "rubber–stamp" their suggestions, if indeed all participants are to share ownership in the final product.

This consultative process does not end with the completion of data analysis. Rather, it is useful to maintain this same level of communication through the phases of dissemination and utilization of results. The efforts of one institution to apply specific findings from the evaluation (e.g., to intensify the delivery of long–acting methods in areas identified by the evaluation to be underserved) may provide an example to other institutions as to how academic research can be applied to improve programs. Moreover, the sense of common purpose developed through the evaluation process may serve to reinforce collaboration in the area of service delivery, even among groups that do not generally work together.

Establishing a Timetable for Specific Activities

The evaluation plan will outline a series of data collection activities corresponding to the objectives of the evaluation, which will be staggered over the life of the project (see Figure III–6). For each type of data, it is essential to specify how often it will collected and reported. (For example, routine service statistics might be collected monthly but reported on a quarterly basis.)

The information for each type of data collection can then be summarized to provide an overview of this activity over the life of the project. This exercise has the dual advantage of indicating (1) possible scheduling conflicts among evaluation activities, especially where data collection personnel are limited, and (2) possible conflicts with other activities of the organization (e.g., preparation of the annual report) or larger community (e.g., major holiday periods when work slows down).

An example of a summary timetable for multiple data collection activities, analysis, and dissemination is shown in Figure V–1.

Budgeting the Evaluation

It is essential to estimate the costs of the proposed evaluation, lest one's plans be stymied by financial realities. Steps in developing the budget include the following:

- Identify the resources for evaluation in the lead organization (and if applicable, in other participating organizations) that are already covered by other sources and will be made available to the activity at no additional cost.[44]

[44] Some organizations may choose to budget these costs, even if they will be covered by existing funding, to be able to track and report the full costs of evaluation for their organization.

Figure V–I

Example of a Timetable for Evaluation Activities

	Year 1				Year 2				Year 3				Year 4				Year 5			
Develop Evaluation Plan	x																			
Steps																				
Clarify objectives of program	x																			
Describe how program should work (conceptual framework)	x																			
Establish objectives of evaluation	x																			
Identify components to be monitored	x																			
Define relevant indicators	x																			
Identify sources of data	x																			
Develop plan for data collection and processing	x																			
Design format for presenting results	x																			
Design/Implement Monitoring Activities																				
Service statistics																				
Review/improve (if necessary) MIS for routine service statistics		x																		
Routinely collect/report service statistics (lg. X denotes annual report)		x	x	X	x	x	x	X	x	x	x	X	x	x	x	X	x	x	x	X
Facility – Based Surveys (illustrative)																				
Example: Situation Analysis (or Service Availability Module)			x	x															x	x
Periodic Qualitative Assessments of Service Delivery (illustrative e.g.)																				
COPE					x	x														
Focus groups among users of adolescent reproductive health services													x	x						
Special Studies of Functional Areas [1] (illustrative)																				
12 month follow up of participants in NORPLANT® training							x	x												
Monitoring the reach and effects of IEC campaign					x	x														
OR project on impact of improved counseling on continuation rates													x	x	x					
Design of Impact Assessment																				
Randomized Experiment (illustrative examples)					x	x	x	x	x	x	x	x	x	x	x	x				
Example: NORPLANT® introduction																				
Multilevel Regression Model (see Chapter IV)																				
DHS and SAM data collection		x	x																	
Follow–up DHS/SAM data collection																	x	x		
Analysis to measure impact																			x	x
Utilize Evaluation Results to Modify/Improve Program		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

[1] Often done to respond to specific concerns that may not be known at onset of programs. Thus, time and funds should be allocated for such activities, but the exact nature and timing of the studies should be determined by actual need.

Examples: Staff salaries and fringe benefits, office space, office equipment such as computers and photocopier, vehicles for field work, etc.

- (If not covered by the above) Estimate the personnel and other direct costs for coordinating the different components of the evaluation.

  Example: The salary and fringe benefits of an individual employed to collect and synthesize service statistics from different branches of the program or from different participating agencies.

- Estimate the costs for each individual data collection activity to be conducted as part of the overall evaluation.

  Example: If the evaluation were to consist of (1) monitoring routinely collected service statistics, (2) conducting a situation analysis at the start and end of the five–year project cycle, (3) conducting a DHS at the start and end of the five–year project cycle,[45] (4) applying the COPE methodology in 30 SDPs, and (5) analyzing the cost per CYP by type of service delivery mechanism, then it would be important to budget each item separately, taking care not to double–count items that would appear in more than one category, such as the purchase of microcomputers with multiple uses.

- Estimate the costs for data processing and analysis.

  Example: In–house personnel may be able to handle the processing and reporting for routinely collected service statistics. By contrast, there are generally substantial costs associated with the processing of situation analysis, DHS data, and multilevel analyses to assess impact.

- Estimate the costs for dissemination of results.

  Example: As will be described in Chapter VI, there are multiple channels for dissemination. Given the importance of this final step of the process, it is essential that funds be budgeted for this purpose from the start.

Budgeting is a skill that may be unfamiliar to some evaluators. Just as one seeks out specialized assistance for sampling, study design, etc., it may also be desirable to seek out the assistance of those with expertise in budgeting in a given institution to bring reality to the budget estimates and ensure that all line items are included.

[45] In the case of Situation Analysis and DHS, the costs will vary greatly by country; budgets need to be developed in conjunction with personnel from these projects with experience in the budgeting process specific to these activities.

Chapter VI

Disseminating and Utilizing the Results

- Tailoring the Presentation of Results to the Intended Audience
- Highlighting the Results that have Important Programmatic Implications
- Establishing a Forum for Presenting Results that Translate to Action
- Remaining Involved in the Process

Chapter VI

Disseminating and Utilizing the Results

TAILORING THE PRESENTATION OF RESULTS TO THE INTENDED AUDIENCE

Top–level policy makers may want the "key findings" from a given evaluation, for example:

- has contraceptive prevalence risen?
- what provinces have the highest rates of contraceptive prevalence?
- what contraceptive methods are most popular?
- is the program reaching the rural area?

Moreover, given the barrage of information that top–level decision makers receive on a daily basis, the report for this audience should be readily comprehensible and aesthetically attractive. For this reason, the medium of choice may be an eye–catching leaflet or thin booklet, containing an executive summary of key results, illustrated with color graphics.

Program administrators, family planning researchers, and others more closely related to the program will want far more detail than contained in an executive summary. This audience will benefit from the full report, with a clear and detailed table of contents, directing the reader to specific topics covered. Graphics should also be used in this type of report, although some readers will also want to see the "numbers behind the graphs." Details of the methodology should be clearly presented, but some of this material may be presented in appendices if beyond the interest or comprehension of the average reader. Where complex multivariate methods are used, the results should be presented in a format that has meaning for those with little or no statistical background.

This full report will generally constitute the most complete version of the evaluation, and it generally appears in the main language of the country.

The evaluation results should be fed back into the system, to program managers and service providers whose work is reflected in the evaluation. Busy practitioners may be put off by a thick volume of research findings, but would be highly receptive to selected findings that relate directly to service delivery (trends in prevalence, changes in method mix, characteristics of users versus non–users, level and geographic concentrations of unmet need, etc.). This group would also benefit from the readily comprehensible brochures for policy makers; another useful vehicle for communicating with this group would be through direct presentation of findings (conference, seminar) followed by a question and answer session and/or small group discussions.

A fourth possible audience is the research community, reachable through international journals and conferences. In the case of a journal article, it is important to focus on a specific topic. Much of the descriptive text that appears in a full country report must be condensed into a few paragraphs that communicate the contribution of this particular study to the scientific literature. Whereas the full report may try to cover most of the variables included in the analysis, a journal article or conference paper will tend to highlight the key points only.

HIGHLIGHTING THE RESULTS THAT HAVE IMPORTANT PROGRAMMATIC IMPLICATIONS

The first three audiences described above — policy makers, program managers and service delivery personnel — will have one question in common:

[46] The authors wish to acknowledge the contribution of Ms. Ann Laughrin of the Futures Group International for the preparation of this chapter.

"What does this mean to me?" Thus, it is essential for the evaluator not only to present the results, but also to interpret their relevance to ongoing programs. In many cases, the audiences in question will grasp the program implications without their being clearly articulated (e.g., contraceptive prevalence is 50% in urban areas, compared to 20% in rural areas). However, most members of the audience will benefit from having the evaluators state the obvious, if only to reinforce their own interpretation of the findings. Moreover, this approach to the presentation of data makes it seem more practical and useful or "less academic."

The fourth audience — the research community — will in many cases also will be interested in the programmatic implications, especially in journals or conferences that are applied in nature.

ESTABLISHING A FORUM FOR PRESENTING RESULTS THAT TRANSLATE TO ACTION

There are a growing number of well–trained program administrators and program managers who can read a report containing statistics and instantly derive what it all means to their program. However, it is common that those in a position to act on the findings will benefit greatly from the opportunity to:

- discuss and better understand the findings;
- internalize this information;
- reconcile this "new information" with their understanding of the program as it currently operates;
- verbalize the implications of the findings in their own terms;
- identify actions that address the situations uncovered by the evaluation; and
- arrive at a plan of action to capitalize on areas of strength and improve on areas of weakness.

The ideal is a forum that promotes discussion and interaction among key individuals in a position to influence future program direction. The process of group dynamics will cause those present to question the results, seek to understand the underlying factors that explain the findings, and through this process, place greater importance on them. However, there may be two exceptions to the value of collective interpretation of the results. First, if the persons present feel defensive about the results, then this type of session may be counter–productive. Second, if any aspects of the data are of questionable validity, then the discussion may focus more on what's wrong with the data than what's wrong with the program.

REMAINING INVOLVED IN THE PROCESS

All too often, evaluators are called in to design and conduct an evaluation; they present their results, conclude with a list of recommendations, and leave. In the ideal scenario, the program administrator or manager will try to incorporate the recommendations in restructuring the program. In many other cases, the old adage holds true: out–of–sight, out–of–mind.

There are three reasons why it is valuable for evaluators to remain in contact with program managers over the period of implementing changes based on the evaluation results.

- The evaluator is available for further consultation regarding the recommendations. For example, where the results of multilevel regression point to expected changes in outcome indicators that would result from implementing certain program actions, the program administrator may seek further clarification on the modifications in service delivery that would be needed to bring about the change.
- Regular contact between the program administrator and evaluator serves as an important reminder that the (original) evaluation was done for a purpose: to identify areas for further improvement. In direct contrast to "out–of–sight, out–of–mind," the presence of the evaluator serves to promote action in the areas identified by the evaluation.
- The evaluator can assist in setting up subsequent designs to determine whether the instituted changes in service delivery in fact bring about the expected changes in service utilization, contraceptive prevalence, etc. As such, this reinforces evaluation as an ongoing process, not a one time event.

Chapter VII

Adaptations to Other Reproductive Health Interventions

- STD/HIV Prevention
- Safe Pregnancy
- Breastfeeding
- Women's Nutrition
- Adolescent Reproductive Health Services

Chapter VII

Adaptations to Other Reproductive Health Interventions

In the wake of the 1994 Cairo International Conference on Population and Development, “family planning” is rapidly expanding from a narrow focus on contraception to a broader range of reproductive health services. As program administrators and service providers move increasingly into the realm of reproductive health, evaluators must address the question: how does one evaluate these interventions?

This chapter focuses on the following areas of reproductive health: STD/HIV prevention, safe pregnancy, breastfeeding, women’s nutrition, and adolescent reproductive health services. “Adolescent services” refers not to a separate health area, but rather to a specific target population. Nonetheless, with the growing recognition of the need for services for this group, it is useful to consider the implications for evaluating such programs.

The guidelines contained in Chapters I–VI of this manual for evaluating family planning programs are generally applicable to other reproductive health interventions. In terms of program–based measures, the indicators may change but the basic approach is similar across other areas of reproductive health. The evaluator is interested in monitoring:

- policy environment;
- the number and types of activities carried out in different functional areas;
- access to services;
- quality of care; and
- level of service utilization (e.g., number of visits, number of new clients, volume of commodities distributed).

By contrast, there are marked differences between family planning and other areas of reproductive health in terms of target population, population–based measures of outcome, and feasibility of data collection. This chapter is organized to cover these issues for each area of reproductive health. It does NOT revisit the issues of demonstrating causality (except in the case of women’s nutrition) discussed in earlier chapters of this manual, which are equally applicable to family planning as well as to other areas of reproductive health.

It is important to recognize that family planning has been one of the most closely evaluated public health interventions in the international health arena. The existing literature reflects over 30 years of concerted effort to find the most methodologically sound yet practical means of evaluating this type of program. By contrast, attempts to evaluate other areas of reproductive health are more recent; much of the work has taken place since the mid–1980s. With greater implementation of reproductive health interventions, evaluators will become more knowledgeable in adapting existing evaluation methodologies to the special circumstances of these other programs.

STD/HIV PREVENTION

Target Population

Reproductive health interventions generally target the population at risk. For certain areas of reproductive health (e.g., family planning, safe pregnancy), this includes all women of relevant age groups in the country. Actual interventions tend to exclude those who by virtue of economic status are able to obtain services through private sources; however, success in a given country is measured in terms of the total population of women — or of married women — in the age range (e.g., contraceptive prevalence, percentage of women delivering under supervised conditions).

With respect to STD/HIV infection, by contrast, not everyone in the general population is at risk. STD interventions target (1)"spread–cluster groups" (those most responsible for maintaining the STD epidemic), (2) symptomatic individuals seeking relief for their symptoms, (3) individuals at high risk for infection due to behavior and/or biologic susceptibility, or (4) those already infected (e.g., unborn children of pregnant women with syphilis). In most countries, it would be a waste of resources to try to reach the general population and a highly ineffective means of reaching those at risk. Indeed, the risk of STD/HIV varies markedly within a given country, region, or even a city. The risk is related to (1) the level of STD/HIV prevalence in a given population or geographical area, and (2) the sexual norms and practices of the sub–populations. STD/HIV interventions tend to be targeted to groups of persons with high risk behaviors: commercial sex workers, truck drivers, migrant workers, and adolescents. Thus, programs need to be evaluated in relation to effects on the behavior of these populations.

In contrast to most other areas of reproductive health that target women of reproductive age, STD/HIV interventions must target both men and women. Some would argue that men are even more important than women, given their role in sexual (and other) decision–making in many countries. DHS–type surveys focus primarily on women of reproductive age, but in recent years have expanded to include males (as an independent sample of randomly selected males or a husband/partner sample). Given that STD/HIV risk behavior appears to be associated with marital or partnership status, it is preferable to collect independent male samples rather than the husband/partner samples in order to generate more valid national estimates of key dependent or independent variables relating to STD/HIV.

In short, there is no standard or conventional target population for STD/HIV interventions. It differs by program or country and must be assessed in each situation.

Outcome Measures and the Feasibility of Data Collection

The primary objective of STD/HIV interventions is to stop the spread of infection, particularly HIV infection. Logically, it would seem that the long–term measure of success should be the level of HIV infection in a given country. Similarly, it would be useful to track changes in the incidence of infection (i.e., rate of new infections).

However, to date it has proven virtually impossible to obtain data on HIV prevalence among a randomly selected sample of the general population in a given country for several reasons.[47] First, testing for HIV requires biologic specimens (i.e., blood, urine, or saliva) that are more difficult to collect than the usual verbal responses in the context of a national survey. Second, the expense of this type of survey is considerable. Third and possibly most important, HIV testing poses several ethical dilemmas: can one measure HIV status and not inform respondents of the results? Is it ethical to inform persons that they are sero–positive but not be able to offer counseling or treatment? How does the program deal with the stigmatization that sero–positive individuals would experience in many societies?

Data are sporadically available from facilities that provide pre–natal and/or obstetric services; indeed, many of the commonly quoted data on HIV prevalence are based on this source of information. Although this might seem to be an easy means of obtaining data on HIV prevalence, in fact it has numerous limitations. First, the women using the services are not necessarily representative of the general population. Second, this type of testing would add an additional procedure for facilities that do not routinely test for HIV and are not necessarily set up to do so. Third, from the perspective of the women in delivery, such testing would mean an additional needle prick. Fourth, given that hospital records are often of poor quality, there is no reason to believe data on HIV status would be better, except in the context of a special research project. In sum, whereas data of the sero–status of women in tertiary facilities is useful for policy purposes (to reflect the magnitude of the problem in general terms), these data are not very satisfactory for evaluation purposes (Hassig, 1995).

[47] There are a few examples of population–based sampling for HIV testing in Africa: in the region of Mwanza, Tanzania (Grosskurth et al., 1995) and in the Rakai district of Uganda (Wawer et al., 1995). However, these are exceptions to the rule.

Finally, while determining HIV prevalence may be the long–term objective, it is not a useful measure for evaluating program impact, as it does not change readily in response to changes in desired practices or behavior. That is, even if it were possible to stop all further transmission of HIV infection, this would not result in an immediate decrease in HIV prevalence. To cite one example based on modeling in a hypothetical high–incidence population, Mertens et al. (1994) estimate that if HIV incidence decreased by 25% over five years due to a successful prevention program, there would be minimal if any decrease in observed HIV prevalence in that same time frame.

In short, sero–prevalence surveys have not been used widely as a means of monitoring the AIDS epidemic and/or evaluating interventions. Even if such data were available, it would be difficult to associate changes in the incidence of HIV infection with program interventions (in the absence of a randomized experiment).

Some have argued that STD prevalence is a useful proxy for HIV infection. Indeed, in Thailand, data on decreases in STD rates are cited as a hopeful sign for decreasing the spread of HIV (Hanenberg et al, 1994). Because there are treatments, if not cures, for most STDs, the ethical dilemmas of testing for STDs are somewhat less than those associated with HIV. However, the technical/operational difficulties are actually more challenging. First, the available screening methods for most STDs (syphilis is the exception) are either expensive or require significant infrastructural support for preservation and transportation of specimens, as well as the preparation and actual testing of samples for presence of infection (e.g., LCR/PCR[48] or culture for herpes or gonorrhea). Second, because several STDs are generally found in varying proportions in populations, measurement of a single STD may be inappropriate; evaluators would need to establish an STD profile at the outset if they were to track a single STD for evaluation purposes. Third, some STD screening tests require confirmatory testing to ascertain the nature (e.g., new, old, treated syphilis) of the infection, while some STDs (e.g., herpes) are chronic conditions (like HIV) and an individual's sero–status will not change even if treatment or behavior change occurs. Fourth, while the existence of therapy/cure options for STDs does make testing for them more theoretically feasible, few countries are well–supplied with sufficient drugs to deal with the problem. Finally, only a few STDs can be identified by relatively easy invasive procedures such as venipuncture or finger prick samples. Most require swabs or other genitally collected specimens, which are not easy to obtain; however, urine–based testing or self–administered swabs offer potential means of making population–based testing more feasible.

Emerging Solutions to these Measurement Problems

There is widespread recognition of the importance of including men in evaluation research on STD/HIV. Indeed, the DHS and CDC Reproductive Health Surveys increasingly include male respondents.

However, the nature and scope of the STD/HIV intervention must be carefully considered when attempting to interpret any data from such surveys for evaluation purposes. For example, a prevention program that focuses on military and university males, migratory laborers such as truckers, and several geographically–defined populations of women at higher risk may have a major impact on the dynamics of the twin epidemics of STD and HIV by changing sexual behavior in those groups, yet yield little if any evidence of change at the population level of the DHS survey.[49] Most STD/HIV prevention programs are neither funded nor implemented at a sufficiently high level (i.e., in terms of intensity and coverage) to have "national" impact.

For programs targeted to a specific sub–population, the sentinel site approach is emerging as a promising evaluation strategy. These pilot data collection efforts have focused on non–representative but programmatically relevant populations. In the above example, a program might set up a "behavioral surveillance survey" in the university and military populations that would

[48] LCR refers to ligase chain reaction; PCR is a polymerase chain reaction. Both are DNA amplification techniques.

[49] Thailand is an exception to this rule. It now appears that if coverage to "spread cluster groups" is high enough, it can have an impact among the general population.

allow the measurement of critical behaviors (and possibly other indicators) at meaningful intervals. For a description of how this approach has been successfully used in Thailand, see AIDSCAP Evaluation Tools, Module 4 (AIDSCAP, 1995).

SAFE PREGNANCY

Target Population

Safe pregnancy refers to improved pregnancy outcomes for women and newborns. The target population for safe pregnancy initiatives consists of two distinct groups: (1) women of reproductive age, and more specifically women who have experienced a pregnancy during a designated period of time, and (2) newborn children (during a designated reference period). Comprehensive program outcome assessment measures thus require "case" information that includes termination of a given pregnancy as well as events occurring during the neonatal period for pregnancies resulting in live births.

At one level, safe pregnancy initiatives involve the entire target population (as defined above) to the extent that "routine" care is advocated by health authorities for all pregnant women and newborns. Such services may be provided by many types of service providers.

Of special concern, however, is the preparedness and performance of health systems and service providers in identifying and managing complications of pregnancy, labor, and delivery, as well as those arising during the newborn/postpartum period. Of particular importance are maternal complications that threaten the lives of pregnant women (i.e., hemorrhage, sepsis, prolonged/obstructed labor, septic abortion, and hypertensive disorders), and conditions that are primary threats to the survival of newborns (i.e. asphyxia, birth injury, sepsis, tetanus, and others mediated through low birth weight). These require specialized care not available locally in many developing country settings, necessitating referral to higher–level facilities and involving more difficult logistics and higher costs for both families and service providers. Thus, cases of complications constitute a "target" of particular interest for evaluating safe pregnancy program initiatives.

Outcome Measures and Feasibility of Data Collection

The primary objectives of safe pregnancy initiatives are to protect the health of pregnant women and newborns by preventing adverse pregnancy outcomes (i.e., miscarriages and stillbirths perinatal, neonatal, and maternal deaths) through timely and effective detection and management of complications arising during pregnancy, delivery, and the neonatal period. Several classes or types of outcome indicators have been identified as being useful for evaluating safe pregnancy interventions.

One group of indicators consists of what may be viewed as "ultimate" outcome measures. Included in this group are several pregnancy–related mortality rates. For example:

- maternal mortality ratio and rate,
- perinatal mortality rate, and
- case–fatality rates (for all complications).

The maternal mortality ratio, which is defined as the number of maternal deaths per 100,000 live births during a specified reference period, provides a measure of obstetric risk once a woman becomes pregnant. The maternal mortality rate, the number of maternal deaths during a given reference period per 100,000 women of reproductive age, provides a broader measure of the likelihood of both becoming pregnant and dying during pregnancy or the puerperium. Where civil registration systems are relatively complete, both the numerator and denominator of the maternal mortality ratio, as well as the numerator of the maternal mortality rate, can be derived from this data source. The denominator of the maternal mortality rate is usually derived from official population projections or estimates. However, since vital statistics systems in most developing countries tend to be incomplete, the numerator data of both measures need to augmented with data from other sources (e.g., hospital death reports, morgue and cemetery reports, etc.). It should be noted that even in developed countries, maternal deaths have been underestimated by up to 50 percent.

The primary alternative measurement approach, the "direct" estimation of maternal mortality in sample surveys, has proven difficult to implement because of the large sample sizes needed to obtain sufficiently precise estimates and the biases associated with households being dissolved as a result of women having died. More promising are "indirect" estimates derived through the "sisterhood method."[50] Although less demanding in terms of sample size than direct estimates, relatively large samples are nevertheless required if geographically disaggregated estimates from the sisterhood method are desired. Another disadvantage is that estimates of the level of maternal mortality produced by the method refer to a point of time in the relatively distant past, usually 10 or more years prior to the survey date.

Similar sources of data and general limitations apply to the perinatal mortality rate (the ratio of late fetal deaths plus early neonatal deaths per 1,000 live births during a designated period of time). However, because of recall problems with dates of late fetal and early infant deaths and serious under–reporting of deaths occurring soon after birth in many settings, it is questionable whether surveys can provide a suitable alternative to vital statistics and/or hospital records.

The case–fatality rate, which is defined as the ratio of the number of deaths from specific complications of pregnancy to the number of complicated obstetric cases presenting at a specific health facility during a designated period of time, measures the likelihood that a woman experiencing an obstetric complication will survive once she enters that health facility for treatment. Aggregated across facilities, it provides a direct measure of the capacity of the health system to deal with obstetric emergencies. However, in order to obtain estimates for the entire population, relatively complete coverage of all facilities offering emergency obstetric services is required.

A second group of indicators pertains to the characteristics of late fetal and infant deaths. For example:

- percent distribution of perinatal deaths by age at time of death;
- ratio of fresh to macerated stillbirths;
- birth weight proportionate mortality rate; and
- birth weight specific mortality rate.

The primary utility of these indicators is to provide information on the relative contributions of failures of different components of the service delivery process (i.e., prenatal care, delivery care, newborn care, infant care) to perinatal mortality and, accordingly, on the component(s) that require further strengthening.

Although both facility and survey data could theoretically be used to measure these indicators, it is questionable whether survey respondents in many developing countries can provide sufficiently accurate information on age/length of gestation at time of death and distinguish fresh from macerated stillbirths. If not, it is necessary to rely upon facility data. Even here, however, there are problems with determining exact ages at time of death. Since most births still are delivered at home in developing countries and birth weights are difficult to measure in the home, facility–based data do not necessarily provide the information needed to determine failures in health services.

The third group of indicators consists of what may be termed "knowledge/coverage" indicators. This group includes a series of typical KAP (e.g., proportion of adults knowledgeable about complications of pregnancy and childbirth, percent with knowledge of location of essential obstetric services, etc.) and "coverage" indicators (e.g., proportion of population residing within one hour's travel time of a facility offering essential obstetric care [EOC], proportion of women attended at least once during pregnancy by trained medical personnel, etc.). In addition, one indicator that directly links supply and demand for obstetric services has been suggested as being especially useful: met need for emergency obstetric care. This indicator has been operationally defined as the proportion of women estimated to have direct obstetric complications seen in a facility that can provide emergency obstetric care during a specified time period.[51]

Most of the knowledge/coverage indicators can be accurately measured on the basis of

[50] The sisterhood method is based upon reports of cases of sisters of survey respondents who died after age 15 of pregnancy–related causes. The method converts this information into an estimate of the maternal mortality ratio. Details on the method may be found in Graham et al., (1990).

[51] See "Indicators for Safe Preganancy," the EVALUATION Project, (1995) for precise definitions of the terms involved in these indicators and a discussion of their uses.

conventional sample surveys. The measurement of access to facilities providing emergency obstetric care is greatly enhanced where Service Availability Modules undertaken in conjunction with DHS and/or Situation Analysis Studies have been conducted, as well as where Geographic Information Systems (GIS) are in place. The demand–related measure, although interpreted on a population basis, requires facility–based data on numbers of complications for the numerators. There is also some question as to the validity of the standard assumption that 15% of all pregnancies will result in complications.[52]

Emerging Solutions to these Measurement Problems

Pending improvements in vital registration and continued refinement of the sisterhood method hold the greatest promise for the measurement of maternal mortality. Recently, small–area censuses and small "coverage–like" surveys have been proposed in order to provide more current and geographically disaggregated estimates than are provided by the sisterhood method. Such efforts tend, however, to be costly.

One of the key issues under study is whether some indicators currently measured on the basis of facility data can be measured with reasonable accuracy in household surveys. Ways to obtain more accurate information on age/length of gestation at time of death and birth weights in household surveys are much needed. The conduct of "verbal autopsies" on larger scale, perhaps in conjunction with large household surveys such as the DHS, holds some promise for obtaining true population–based data on a number of the indicators, but requires further testing in the context of measuring safe pregnancy outcomes.[53] The Subcommittee on Safe Pregnancy of The EVALUATION Project's Reproductive Health Indicators Working Group has also proposed an initiative to revise record–keeping at health facilities on cases of pregnancy complications in order to improve the quality of information available for program monitoring and management.

BREASTFEEDING

Target Population

The target population for breastfeeding programs includes all fertile women of reproductive age who are at risk of conception in a given population. Particular emphasis is placed on women who are pregnant, in the perinatal period, or lactating. There also are sub–groups with special needs: first time mothers, working mothers, and women with previous breastfeeding "failure." In addition to the specific groups, certain educational or promotional interventions also target entire populations to foster a culture supportive of breastfeeding in homes, communities, workplaces, and health services.

National breastfeeding programs are generally evaluated in reference to the breastfeeding practices among women having given birth recently. By contrast, interventions targeting specific subgroups should be evaluated in reference to members of those subgroups. For example, educational interventions aimed at adolescent and school–aged children could be assessed in terms of changes in knowledge and attitudes among the group.

Outcome Measures and Feasibility of Data Collection

The most recent policy on duration of breastfeeding (UNICEF/UNESCO/WHO/UNFPA, 1994) states that all infants should be fed exclusively on breast milk from birth to 6 months. The World Health Assembly Confirmation of the marketing code for breastmilk substitutes states that supplementary foods should be introduced at about six months.

Breastfeeding programs are generally designed to increase (1) the incidence of breastfeeding, (2) the prevalence of exclusive breastfeeding in the first 6 months of life, and (3) the duration of breastfeeding. Specific interventions may target a particular obstacle to optimal breastfeeding practices, such as (1) lack of knowledge among

[52] See "Indicators for Safe Pregnancy," The EVALUATION Project (1995), and World Health Organization (1993), "Indicators to Monitor Maternal Health Goals", for further discussion of this issue.

[53] "Verbal autopsies" are survey instruments used to gather information on signs and symptoms of illness that preceded the death of a household member reported as having died. Details on the method may be found in Gray et al. (1990) and Kalter et al. (1990).

mothers and health care providers about the health, nutrition, economic, and child spacing benefits of breastfeeding, (2) birthing and health care protocols which interfere with successful lactation, or (3) absence of lactation management expertise in communities and health services that can assist mothers in successfully initiating breastfeeding at birth. Communication/education programs emphasize avoidance of breastmilk substitutes, delay of introduction of supplementary fluids and foods, and ways to resolve lactation problems should they occur.

Outcome measures for breastfeeding program evaluation frequently include initiation of breastfeeding in the first hour of life, duration of amenorrhea, use of lactational amenorrhea for contraception, exclusive breastfeeding rates, mean duration of breastfeeding, and frequency of feeds.

The preferred sources of data for measuring outcomes related to breastfeeding are the DHS–type survey and simplified cluster survey approaches. In most countries, it is relatively easy to ask questions about breastfeeding in a private interview situation. The greatest measurement problem is recall bias. It may be difficult for women to accurately recall the exact number of feeds given to the baby, even in the previous 24 hours, or the exact age in months when breastfeeding stopped. Moreover, the researcher's definition of exclusive breastfeeding may differ from the mother's, and thus indicators need to be carefully operationalized for breastfeeding. However, in comparison to other areas of reproductive health, the outcome indicators can be measured fairly easily and reliably using sample surveys.

If DHS surveys are not available as a source of data for outcome assessment of national programs, the preferred alternative would be a survey among a large, representative sample of women having recently given birth (since recall of breastfeeding practices in the past is notoriously unreliable). In designing such surveys, it is important to conduct preliminary qualitative research on the meanings of terms (e.g., exclusive breastfeeding) and to carefully pretest the instruments. However, the cost of such surveys is high, putting this option beyond the possibilities of most programs.

Although the focus of this manual is on national programs, many breastfeeding programs are sub–national in scope, operating in selected service delivery posts or communities. Evaluation of such programs would not be feasible utilizing national level survey data. If a facility–based or community–based evaluation is to be undertaken, however, it is important to clarify (1) what the specific outcomes/objectives of the program are (e.g., initiation within one hour of birth, exclusive breastfeeding for six months), (2) who will use the information generated, and (3) what purpose the evaluation is to serve. Typical users are program implementers, program funders, and policy makers.

One source of data for obtaining information about neonatal breastfeeding practices (e.g., initiation of breastfeeding during the first hour of life) is the hospital/maternity–based survey or exit interview. Exit interviews may be particularly useful if the target group of the intervention is women who give birth in health facilities. It also is a good strategy for assessing the extent to which hospitals are effectively executing baby–friendly initiatives.

Emerging Solutions to these Measurement Problems

For evaluation purposes, program administrators and funding agencies would ideally like to know the effect of their interventions on the breastfeeding practices of the target population. The ever–growing data bank of DHS surveys provides a promising opportunity for further empirical evaluation of breastfeeding practices at the national or regional level.

Researchers in turn are interested in demonstrating the effect of infant feeding practices on fertility, infant nutrition, infant health, and related outcomes. The latter requires large samples and preferably longitudinal studies that are challenging to conduct and analyze (since the impact of infant feeding practices is also affected by infant's age, mother's health, and the general socio–economic conditions of the household).

To date, the evaluation research in this area has treated these two issues separately: (1) the effects of program interventions on breastfeeding practices, and (2) the effects of breastfeeding on fertility and mortality. The challenge remains to demonstrate empirically in the context of a given study that program interventions not only affect behavior in the medium term but also influence fertility and health status in the long run.

Whereas demonstrating impact remains one objective in the evaluation of breastfeeding interventions, there is also much to be learned from less rigorous methods, which are useful in improving programs at the local level. For example, neonatal and infant health outcomes tracked before and after “baby–friendly” hospital interventions can provide a useful means of assessing their effectiveness among select populations.

WOMEN’S NUTRITION

Target Population

The definition of the target population for women’s nutrition interventions in the context of reproductive health is complicated by the cumulative effects of malnutrition. Poor nutrition during early childhood can and does cause irreversible physical effects. Growth deficits caused in part by malnutrition during childhood are never recovered by many women in the developing world, resulting in shortness of stature, cephalopelvic disproportion, poor pregnancy outcomes, and possibly other physiologic consequences. Therefore, interventions targeted at women’s nutrition problems can potentially cover a broad age spectrum.

The most common forms of malnutrition in the developing world are protein–energy malnutrition and micronutrient deficiencies (vitamin A, iodine, and iron). All of these deficiencies are linked to poverty and illiteracy. However, iodine deficiency may be more geographically defined. Vitamin A deficiency also is determined in part by ecologic factors and traditional dietary habits. Iron deficiency is widespread throughout the developed and developing world.

Most of the resources devoted to women’s nutrition, however, are concentrated on pregnant and lactating women for two main reasons: (1) during pregnancy and lactation, nutritional requirements increase, and (2) maternal nutritional deficiencies during pregnancy can have serious consequences for women and their infants.

Adolescent mothers represent a special target population because they must support not only their own nutritional requirements for growth but their fetus/infant’s as well. Where resources permit, however, women’s nutrition programs increasingly target a broader age range. This is being advocated in part because pre–pregnancy nutritional status is an important determinant of reproductive outcomes, and because women’s health issues are now considered as a priority in their own right.

Outcome Measures and the Feasibility of Data Collection

The long–term objective of nutritional programs is to reduce the incidence and prevalence of protein–energy and micronutrient deficiencies. This is, however, an ambitious undertaking. The immediate causes of nutritional deficiencies are inadequate dietary intake and/or poor utilization of nutrients (often the result of infections and poor health care). Nutritional requirements of women in the developing world also are often elevated by high physical workloads and frequent recurring cycles of pregnancy and breastfeeding.

Protein–energy malnutrition (PEM) is most frequently assessed using anthropometric measures such as weight, height and arm circumference, measures that are now routinely collected as a part of the DHS. Shortness among women(<145 cm) reflects nutritional deficiency during childhood and is therefore unlikely to change during the life of most nutritional programs (except possibly in the case of adolescents). Measures such as weight, weight in relation to height, and arm circumference reflect a women’s thinness. These measures should be sensitive to nutritional interventions that are targeted to the PEM problem. These maternal anthropometric measures are easily collected, and indicators based on these are widely viewed as valid and reliable.

The assessment of micronutrient status is more difficult than anthropometric status and thus is not a routine part of DHS. Laboratory assays and/or somewhat invasive techniques are required, as micronutrient status is based on biochemical analysis of blood (iron and vitamin A), breast milk (vitamin A) or urine (iodine). Despite technological improvements, these tests still substantially complicate the logistics of field work. Technological advancements are increasingly simplifying the process and have greatly reduced the financial costs of testing. In some cases, countries have elected to include micronutrient assessment in their DHS programs.

For selected micronutrient interventions such as vitamin A and iodine supplementation and fortification, coverage levels are reasonable proxy measures of intermediate outcome (conceptually similar to immunization coverage). The evaluation of iron supplementation programs and programs targeting PEM cannot rely on proxies at this time.

A problem that confronts all reproductive health interventions is to demonstrate that a given intervention is responsible for change in the target population (indeed, it is a major theme of this manual). Nowhere is this challenge more evident than for women's nutrition programs (especially those targeting PEM and anemia). In part this is due to the multiple underlying causes of malnutrition. Programs frequently do not result in change in nutritional status because (1) they do not target those most in need, (2) they do not target the multiple (dietary and non–dietary) causes of malnutrition, or (3) they do so inadequately.

Household food security and availability, favorable economic conditions, access to health services, and a healthy environment are all necessary conditions for adequate dietary intake and control of diseases. It is most often the poor who are adversely affected by inadequate food availability. The multi–causal nature of poor nutrition in women makes causal attribution particularly complex.

Emerging Solutions to the Measurement Problems

The emphasis in the international health field on micronutrient deficiencies is likely to result in increased availability of data for identifying proxy indicators of micronutrient status. Also, technologic advancements in physiologic/ biochemical assessment techniques will increase the practicality of utilizing these approaches.

ADOLESCENT REPRODUCTIVE HEALTH SERVICES

Adolescent programs are relatively new, and to date there has been little systematic evaluation of their effectiveness in developing countries (Senderowitz, 1995). The limited evidence from developing countries has yet to demonstrate that teen programs reduce sexual risk–taking, pregnancy, and STD transmission (Kirby, 1994). Even in the U.S., empirical evidence on the effectiveness of adolescent programs and how they work is limited (Brown and Eisenberg, 1995).

The recent increase in interest in adolescent programs coincides with a renewed focus on accountability and results in the international reproductive health circles. As these types of programs continue to develop, they will provide for opportunities to evaluate what works and in what contexts.

Target Population

"Adolescent reproductive health services" overlap with the other topics presented in this chapter, since they may include one or more of the areas (STD/HIV, safe pregnancy, breastfeeding, and nutrition) in addition to other social services (counseling, drug prevention, job training) and recreational activities. The common and defining characteristic of this set of multi–faceted interventions is the age of the population they target; yet this age criterion is by no means standard across programs.

Adolescence encompasses physical and emotional stages of transition from childhood to adulthood. Physiologically, adolescence is a period of rapid growth and development of secondary sexual characteristics. It is also a period of emotional turbulence during which the adolescent strives to achieve independence from his parents or guardians. While these stages are universal, they can occur at widely varying ages in different cultures. For this reason, no single, generalizable age criterion has emerged for use in different settings.

Some programs, especially those concerned with contraception, use the 15–19 age bracket, which has the advantage of consistency with DHS–type surveys. However, many programs attempt to reach young people, at least with information and values clarification, long before they become sexually active. Thus, the age range for certain interventions may be as low as 10–12 years. Because the problems of adolescence often continue past the teen years, certain programs continue to extend services to young people into their early twenties (Stewart and Eckert, 1995).

The World Health Organization (WHO) has defined adolescence in two stages. All persons between the ages of 10 and 19 are defined as adolescents, with the younger group from 10 to 14 classified as "early adolescence" and 15 to 19 as "late adolescence." The latter category may be further subdivided into 15–17 and 18–19 brackets, where programmatically appropriate. WHO further suggests that the terms "youth" may be used to refer to persons between the ages of 15–24, and "young people" for the entire age group of 10–24 (WHO, 1989).

A second possible dimension used in defining the target population for adolescent programs is marital status. In many countries adolescent programs have evolved to meet the reproductive health needs of unmarried young women who do not feel comfortable using the same services as older, (mostly) married women. This focus is particularly appropriate where adolescents become sexually active at a relatively early age and/or marriage is postponed for educational or other reasons. However, in parts of South Asia and the Middle East, there is a need to focus services on young married couples, especially where existing programs ignore the needs of nulliparous women or even make childbirth a condition for getting family planning services (Mensch, 1995).

In short, there is no standard definition of adolescence. The target population used to evaluate adolescent reproductive health interventions should be consistent with the criteria established by the program in question.

Outcome Measures and the Feasibility of Data Collection

Objectives

Adolescent reproductive health programs have multiple objectives that differ from one program to another. For example, some programs are designed primarily as educational interventions to increase knowledge, create awareness, and form positive attitudes; they may have no service component. Others, such as comprehensive health services, may provide multiple services (e.g., family planning, STD treatment, nutritional counseling, shelter from abusive situations, drug counseling, etc.). Still others may offer one or more non–health services, such as income–generating activities, legal services, employment counseling, recreational activities, and so forth. Anecdotal evidence suggests that programs with a broad range of activities may be less controversial than those offering reproductive health services only.

In short, there is no single objective or even set of objectives against which to systematically evaluate adolescent reproductive health services. Rather, the objectives must be defined in relation to specific programs. Moreover, to the extent that reproductive health services are provided in connection with other activities, the evaluation of such programs may also need to track non–health results.

Program– Versus Population–based Measures

If adolescent services were as well developed and far–reaching as are contraceptive services for adult women in many countries of the world, one would expect changes in these population level indicators as a result of program interventions. The reality, however, is that most adolescent programs (where they exist at all) are still in their infancy. They reach a relatively small segment of the population, often limited to the major urban area(s) of the country. While these programs may have a pronounced effect on the individuals they do reach, at present the coverage of such programs is extremely limited.[54] Thus, whereas one would expect to find changes in knowledge, attitudes, and behaviors among the clients or participants in such programs, it is unlikely that adolescent programs in their current form are of sufficient magnitude or intensity to affect population–based measures of behavior.

Conceptually, changes at the population level are the long–term goal of most adolescent programs; however, evaluating such programs in terms of changes at the population level would be of questionable utility. In fact, it could even be detrimental to continued support of such efforts, if those without a full understanding of the technical issues were to conclude that adolescent

[54] One notable exception is the musical video featuring Tatiana and Johnnie that swept through Latin America in the late 1980s; this could aptly be described as an intervention, but not a program *per se*.

programs "don't produce results." In many cases it may be more productive to evaluate adolescent reproductive health services in terms of changes in knowledge, skills, and behaviors among the clients or participants in these programs (Stewart and Eckert, 1995).

Feasibility of Data Collection

Most data collection to date has involved interviews with adolescents (at school, at home, in clinics, etc.), self–administered questionnaires or "tests" (e.g., pretest–posttest to measure knowledge gain from an educational intervention), focus groups, observation in clinical settings, analysis of clinical records, and other standard methods. Special attention must be given to the wording of questions (to be sure that key concepts are presented in the vernacular of adolescents, rather than the terminology of the medical community). Also, special authorization may be required, for example, to administer a questionnaire to adolescents in the school or interview an unmarried teen in her home.

Intermediate outcomes or the behavioral measures of key interest in adolescent programs (e.g., age at first intercourse, use of contraception at first and at last intercourse, unintended pregnancy, self–report of STDs, self–report of drug use, etc.) can be obtained through self–report on surveys. Although there are some issues relating to the validity of the responses given by adolescents to sensitive questions,[55] nonetheless this information can be collected through direct interview or self–administered questionnaire.

The feasibility of measuring long–term outcomes varies by reproductive health area, as reflected in the previous sections of this chapter. For example, it is fairly easy to measure age –specific fertility rates or wanted fertility rates among adolescents (if defined as aged 15–19) based on self–report from a DHS–type study. By contrast, it is highly problematic to measure the maternal mortality rate or HIV prevalence among the adolescent population. Nutritional status is easy to measure, but attribution of change to specific program interventions is difficult.

Emerging Solutions to the Measurement Problems

An important first step in evaluating adolescent reproductive health programs is to develop an inventory of the different objectives common to these programs and to define the corresponding intermediate outcomes. Although it may be possible to measure change at the population level in relatively few cases, nonetheless this inventory would represent a useful menu for those involved in evaluating such programs.

Depending on available resources, programs may take one of two (or both) directions: (1) to obtain descriptive data—often of a qualitative nature—on how the program has functioned and where improvements need to be made, and (2) to conduct experiments or quasi–experiments to measure the extent to which a given intervention affects behavior (either among program participants or in special circumstances at the population level). The former will be important in providing feedback to managers about changes they can make in the short–term and at the local level to increase the acceptability of the interventions to clients, potential clients, and the community at large. The latter will be extremely important in demonstrating to the donor community and other interested parties the type and magnitude of change that can be expected from different types of interventions.

To conclude, the evaluation of adolescent reproductive health interventions is still at such an early stage that it is premature to identify "emerging solutions." Rather, what is promising is the mounting interest in systematically evaluating interventions in this area.

[55] Anecdotal evidence suggests that adolescents seem relatively willing to discuss these subjects frankly if the interviewer is of the same sex and also young (under 30) (Morris, 1995).

Chapter VIII

Summary: Checklist of Steps in Designing an Evaluation Plan

- Define the Scope of the Evaluation (Chapter II)
- Define the Methodological Approach: Program Monitoring (Chapter III)
- Define the Methodological Approach: Impact Assessment (Chapter IV)
- Develop an Implementation Plan (Chapter V)
- Disseminate and Utilize the Results (Chapter VI)

SUMMARY: CHECKLIST OF STEPS IN DESIGNING AN EVALUATION PLAN

This manual is designed to assist professionals in the international family planning/reproductive health community to:

- Differentiate between the main types of program evaluation: program monitoring and impact assessment;
- Critically evaluate the strengths and limitations of alternative methods for impact assessment;
- Assess and select the type(s) of evaluation most appropriate to a given setting;
- Identify appropriate indicators and sources of data for the evaluation; and
- Design an evaluation plan outlining study design(s), indicators, and sources of data that serves as a plan of action for subsequent implementation.

The manual reviews a series of steps to cover in designing an evaluation plan. It assumes that the evaluation will include some type of program monitoring. Where the objective is also to assess impact, many of these same steps apply, but the appropriate study design must be identified and implemented, as outlined in Chapter IV.

Although the original mandate of The EVALUATION Project was to evaluate the impact of family planning programs on fertility, the scope of many family planning programs has expanded in recent years to include other areas of reproductive health. To this end, this manual reviews a series of methodological issues (e.g., target population, measurement of outcome indicators, feasibility of data collection) for evaluating programs in other areas of reproductive health (as described in Chapter VII):

- STD/HIV prevention
- Safe pregnancy
- Breastfeeding
- Women’s nutrition
- Adolescent reproductive health services

The basic steps to follow for program monitoring and impact assessment can be summarized in a checklist format, as outlined in the boxes on page 94.

Checklist of Steps in Evaluating a Program

Define the Scope of the Evaluation (Chapter II)

❐ Determine the program goals and objectives

❐ Describe how the program “should” work (conceptual model)

❐ Establish the objectives of the evaluation

❐ Outline the scope of the evaluation in the evaluation plan

Define the Methodological Approach: Program Monitoring (Chapter III)

❐ Clarify the primary purpose of monitoring

❐ Identify the components to be monitored

❐ Define relevant indicators

❐ Identify sources of data

❐ Design a format for the presentation of results

❐ Summarize the methodological approach

Define the Methodological Approach: Impact Assessment (Chapter IV)

❐ Review the methodological requirements for the “preferred” approaches to assessing impact in family planning programs:

- Randomized experiments
- Multilevel regression methods
- Quasi–experiments

❐ Assess the feasibility of using one of the preferred approaches

❐ (If one of the preferred approaches is possible) Identify and negotiate the special data needs for the specific country setting

❐ (If the preferred approaches are not feasible) Review the alternative approaches to measuring impact:

- Decomposition (Proximate Determinants Model)
- Prevalence Method

❐ Determine and implement the optimal design for the country–specific circumstances.

Develop an Implementation Plan (Chapter V)

❐ Define the institutions and individuals responsible for the evaluation

❐ Establish a timetable for specific activities

❐ Budget for the evaluation

Disseminate and Utilize the Results (Chapter VI)

❐ Tailor the presentation of results to the intended audience

❐ Highlight the results that have important programmatic implications

❐ Establish a forum for presenting results

❐ Remain involved in the process

REFERENCES AND APPENDIX

AIDSCAP, 1995. “Application of a Behavioral Surveillance Survey Tool,” AIDSCAP Evaluation Tools, Module 4. Family Health International, Arlington, VA.

Angeles, G., T.A. Mroz, and D.K. Guilkey. 1995. “Purposive Program Placement and the Estimation of Program Effects: The Impact of Family Planning Programs in Tunisia.” Paper presented at the Annual Meeting of the Population Association of America, San Francisco, CA.

AVSC International. 1995. *COPE: Client–Oriented Provider Efficient Services. A Process and Tools for Quality Improvement in Family Planning and Other Reproductive Health Services*. New York, New York.

Bauman, K., C. Viadro, and A. Tsui. 1994. “Use of True Experimental Designs for Family Planning Program Evaluation: Merits, Problems, and Solutions.” *International Family Planning Perspectives* 20(3): 108–113.

Bertrand, J.T., R.J. Magnani, and J.C. Knowles. 1994. *Handbook of Indicators for Family Planning Program Evaluation*. Chapel Hill, NC: The EVALUATION Project.

Bertrand, J.T., R. Santiso, S.H. Linder, and M.A. Pineda. 1987. “Evaluation of a Communications Program to Increase Adoption of Vasectomy in Guatemala.” *Studies in Family Planning* 18(6):361–370.

Bertrand, J.T. and A. Tsui (eds.), 1995. *Indicators for Reproductive Health Program Evaluation*. Chapel Hill, NC: The EVALUATION Project.

Bogue, D.J. 1970. *Family Planning Improvement through Evaluation: A Manual of Basic Principles*. Community and Family Study Center, University of Chicago, Manual No. 1.

Bollen, K.A., D.K. Guilkey, and T.A. Mroz. 1995. “Binary Outcomes and Endogenous Explanatory Variables: Tests and Solutions with an Application to the Demand for Contraceptive Use in Tunisia.” *Demography* 32(1):111–31.

Bollen, K.A., D. Guilkey, and T.A. Mroz. 1992. “Methods for Evaluating the Impact of Family Planning Programs in Structural Models.” Chapel Hill, NC: The EVALUATION Project.

Bongaarts, J. 1993. “The Fertility Impact of Family Planning Programs,” New York: Population Council, Working Paper No. 47.

Bongaarts, J. 1978. “A Framework for Analyzing the Proximate Determinants of Fertility.” *Population and Development Review*, 4(1): 105–132.

Bongaarts, J. 1986. “The Prevalence Method.” in United Nations *Manual IX Addendum: The Methodology of Measuring the Impact of Family Planning Programmes on Fertility*. New York: Department of International Economic and Social Affairs, pp. 9–14.

Bongaarts, J. and S. Kirmeyer. 1980. “Estimating the Impact of Contraceptive Prevalence on Fertility: Aggregate and Age–Specific Versions of a Model.” in Hermalin, A. and B. Entwisle (eds.) *The Role of Surveys in the Analysis of Family Planning Programs*. Liege, Belgium: Ordina Editions, pp. 381–408.

Bongaarts, J. and R. Potter. 1983. *Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants*. New York: Academic Press.

Brown, S.S. and L. Eisenberg (eds). 1995. *The Best Intentions: Unintended Pregnancy and the Well–being of Children and Families*. Washington: National Academy Press.

Buckner, B., A.O. Tsui, K. McKaig, and A.I. Hermalin. 1995, *A Guide to Methods of Family Planning Evaluation: 1965–1990*. Chapel Hill, NC: The EVALUATION Project.

Campbell, D. and J. Stanley. 1963. *Experimental and Quasi–Experimental Designs in Research*. Boston, MA: Houghton Mifflin Co.

Casterline, J., L. Domingo, and Z. Zablan. 1988. “Trends in Fertility in the Philippines: An Integrated Analysis of Four National Surveys.” Manila, Philippines: University of the Philippines Population Institute.

Chamratrithirong, A., P. Prasartkul, and A. Bennett. 1986. “Multivariate Areal Analysis of the Efficiency of the Family Planning Programme and Its Impact on Fertility in Thailand.” Bangkok: Economic and Social Commission for Asia and the Pacific, Asia Population Studies Series No. 68.

Chandrasekaran, C. and A. I. Hermalin. 1985. *Measuring the Effects of Family Planning Programs on Fertility*. Dolhain, Belgium: Ordina Editions.

Cook, T. and D. Campbell. 1979. *Quasi–Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Co.

Entwisle, B., A.I. Hermalin, P. Kamnuansilpa, and A. Chamratrithirong. 1984. “A Multi–Level Model of Family Planning Availability and Contraceptive Use in Rural Thailand.” *Demography* 21(4): 559–74.

Fisher, A., J. Laing, J. Stoeckel, and J. Townsend. 1991. *Handbook for Family Planning Operations Research*. New York: Population Council.

Fisher, A., B. Mensch, R. Miller, S. Askew, A. Jain, C. Ndeti, L. Ndhlovu, and P. Tapsoba. 1992. *Guidelines and Instruments for a Family Planning Situation Analysis Study*. New York: The Population Council.

Freedman, R. and J.Y. Takeshita. 1969. *Family Planning in Taiwan*. Princeton, NJ: Princeton University Press.

Garate, M.R., et. al., “Comparison Between Two Payment Models to Physicians in Two Private Family Planning Agencies in Peru: Final Report,” The Population Council, Lima, Dec. 1993.

Garcia–Nuñez, J. 1992. *Improving Family Planning Evaluation: a Step–by–step Guide for Managers and Evaluators*. West Hartford, Connecticut: Kumarian Press, Inc.

Gertler, P. and J. Molyneaux. 1994. “How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility.” *Demography*, 31(1): 33–63.

Graham, W., W. Brass, and R.W. Snow. 1989. “Estimating Maternal Mortality.” *International Family Planning Perspectives*, 20(3):125–35.

Gray, R.H., H.D. Kalter, and P. Barass. 1990. “The Use of Verbal Autopsy Methods to Determine Selected Causes of Death in Children.” Occasional Paper No. 10. Baltimore, MD: Institute for International Programs, Johns Hopkins University.

Grosskurth, H., F. Mosha, J. Todd, et al. 1995. “Impact of Improved Treatment of Sexually Transmitted Diseases on HIV Infection in Rural Tanzania: Randomized Control Trial.” *Lancet*: 346:530–536.

Guilkey, D. and S. Cochrane. 1994. “Zimbabwe: Determinants of Contraceptive Use at the Leading Edge of Fertility Transition in Sub–Saharan Africa.” Chapel Hill, NC: Carolina Population Center.

Hanenberg, R.S., W. Rojanapithayakorn, P. Kunasol,and D.C. Sokal. 1994. “Impact of Thailand’s HIV–control Program as Indicated by the Decline of Sexually Transmitted Diseases,” *Lancet*, 344:243–245.

Hassig, S.E. 1995. Personal communication.

Hermalin, A. 1979. “Multivariate Areal Analysis.” in United Nations *Manual IX: The Methodology of Measuring the Impact of Family Planning Programmes on Fertility*. New York: Department of International Economic and Social Affairs, pp. 97–111.

Hermalin, A. 1975. “Regression Analysis of Areal Data.” in Chandrasekaran, C. and A. Hermalin (eds.), *Measuring the Effect of Family Planning Programs on Fertility*. Dolhain, Belgium: Ordina Editions, pp. 245–99.

Hermalin, A. 1982. “Some Cautions in the Use and Interpretation of Regression Analysis for the Evaluation of Family Planning Programs.” in United Nations *Evaluation of the Impact of Family Planning Programmes on Fertility: Sources of Variance*. New York: Department of International Economic and Social Affairs, pp. 265–67.

Jain, A. and J. Bruce, 1994. “A Reproductive Health Approach to the Objectives and Assessment of Family Planning Programs,” in Sen, G. et al. (eds.), *Population Policies Reconsidered: Health, Empowerment, and Rights*, Boston, MA: Harvard University Press. pp. 103–209.

Janowitz, B. and J. H. Bratt. 1992. “Costs of Family Planning Services: A Critique of the Literature.” *International Family Planning Perspectives*, 18: 137–144.

Janowitz, B. and J.H. Bratt. 1994. *Methods for Costing Family Planning Services*, New York: UNFPA, and Research Triangle Park, NC: Family Health International.

Jolly, C. and J. Gribble. 1993. “The Proximate Determinants of Fertility.” in Foote, K., K. Hill, and L. Martin (eds.) *Demographic Change in Sub–Saharan Africa*. Washington, DC: National Academy Press.

Kalter, H.D., R.H. Gray, R.E. Black, and S.A. Gultiano. 1990. “Validation of Postmortem Interviews to Ascertain Selected Causes of Death in Children.” *International Journal of Epidemiology*, 19:380–6.

Koblinsky, M., K. McLaurin, P. Russell–Brown, P. Gorbach (eds.), “Final Report of the Subcommittee on Safe Pregnancy.” 1995. *Indicators for Reproductive Health Program Evaluation*. Chapel Hill, NC: The EVALUATION Project.

Lloyd, C. and J. Ross. 1989. "Methods for Measuring the Fertility Impact of Family Planning Programs: The Experience of the Last Decade." Research Division Working Papers, No. 7. NY: The Population Council.

McInerney, M., and C. de la Quintana, "A Comparative Study of Three Strategies to Improve the Sustainability of a Bolivian Family Planning Provider," The Population Council, La Paz, 1994.

Mensch, B. 1995. Personal communication.

Mensch, B., A. Jain, et al. 1994. "Assessing the Impact of Family Planning Services on Contraceptive Use in Peru: A Case Study Linking Situation Analysis Data to the DHS." Paper presented at the 1994 Annual Meeting of the Population Association of America, Miami, FL.

Mertens, T., M. Carael, P. Sato, J. Cleland, H. Ward, and G.D. Smith. 1994. "Prevention Indicators for Evaluating the Progress of National AIDS Programmes." *AIDS* 8:1359–1369.

Miller, R., K. Miller, L. Ndhlovu, J. Solo, and O. Achola. 1996. "A Comparison of the 1995 and 1989 Kenya Situation Analysis Study Findings." New York: The Population Council (unpublished manuscript).

Morris, L. Personal communication 1995.

Mroz, T.A., and D.K. Guilkey. 1992. "Discrete Factor Approximations for Use in Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variables." Chapel Hill, NC: The EVALUATION Project.

Newman, J. 1988. "A Stochastic Dynamic Model of Fertility." In Schultz, T.P. *Research in Population Economics*. New York: JAI Press, Inc., pp. 41–68.

Ojeda, G., R. Murad, F. Leon, "Testing Pricing/Payment Systems to Improve Access and Cost–Recovery from Norplant®: Final Report," The Population Council, Lima, May, 1994.

Phillips, J.F., W.S. Stinson, S. Bhatia, M. Rahman, and J. Chakraboty. 1982. "The Demographic Impact of the Family Planning–Health Services Project in Matlab, Bangladesh." *Studies in Family Planning* 13(5):131–140.

Poston, D., and B. Chu. 1987. "Socioeconomic Development, Family Planning, and Fertility in China." *Demography*. 24(4): 531–51.

Reinis, K. 1992. "The Impact of Proximate Determinants of Fertility: Evaluating Bongaarts's and Hobcraft and Little's Methods of Estimation." *Population Studies*, 46(2): 309–326.

Reynolds, J. 1993. *Cost Analysis, Primary Health Care Management Advancement Programme, Module 8, Users Guide*. Washington, DC: Aga Khan Foundation and University Research Corporation.

Reynolds, J. 1970. A *Framework for the Design of Family Planning Program Evaluation.* International Institute for the Study of Human Reproduction: New York.

Reynolds, J. and K. C. Gaspari. 1985. *Cost–effective Analysis*. Chevy Chase: Primary Health Care Operations Research Project (PRICOR).

Robey, B., S.O. Rutstein, L. Morris, and R. Blackburn. 1992. “The Reproductive Revolution: New Survey Findings.” *Population Reports*, Series M. No.11.

Rossi, P.H., and H. Freeman 1993. *Evaluation: A Systematic Approach*. Newbury Park, CA: Sage Publications.

Rossi, P.H., J.D. Wright, and A.B. Anderson. 1983. *Handbook of Survey Research*. New York: Academic Press.

Senderowitz, J. 1995. *Adolescent Health: Reassessing the Passage of Adulthood*. Bank Discussion Paper. World Bank, Washington, D.C.

Sherris, J. D., K. A. London, S. H. Moore, J. H. Pile and W. B. Watson, 1985. “The Impact of Family Planning Programs on Fertility.” *Population Reports*, XIII, 1:J733–J771.

Stewart, L. and E. Eckert (eds.), 1995, “Indicators for Adolescent Reproductive Health Services,” in Tsui, A. and J. Bertrand (eds.), *Indicators for Reproductive Health Program Evaluation*. Chapel Hill, NC: The EVALUATION Project.

Suarez, E., and C. Brambila, “Cost Analysis of Family Planning Services in Private Family Planning Programs, FEMAP, Mexico: Final Report,” The Population Council, Mexico City, June 1994.

Tsui, A.O. and P.D. Gorbach, forthcoming in 1996. *Framing Family Planning Program Evaluation: Cause, Logic and Action*. Chapel Hill, NC: The EVALUATION Project.

UNICEF/UNESCO/WHO/UNFPA. *Facts for Life*. New York, NY.

United Nations. 1982. *Evaluation of the Impact of Family Planning Programmes on Fertility: Sources of Variance*. New York: Department of International Economic and Social Affairs.

United Nations. 1986. *Manual IX Addendum: The Methodology of Measuring the Impact of Family* Planning Programmes on Fertility. New York: Department of International Economic and Social Affairs.

United Nations. 1979. *Manual IX: The Methodology of Measuring the Impact of Family Planning Programmes on Fertility*. New York: Department of International Economic and Social Affairs.

United Nations. 1985. *Studies to Enhance the Evaluation of Family Planning Programmes*. New York: Department of International Economic and Social Affairs.

USAID (United States Agency for International Development). 1995. “The Agency’s Strategic Framework and Indicators 1995/96,” Performance Measurement and Evaluation Division, Center for Development Information and Evaluation, Bureau for Policy and Program Coordination.

Veney, J.E. and P. Gorbach, 1993. “Definitions for Program Evaluation Terms.” Chapel Hill, N.C.: EVALUATION Project (Working Paper Series No. WP–TR–01).

Vian, T. 1993. “Analyzing Costs for Management Decisions,” *Family Planning Manager* 2(2):1–18.

Wawer, M.J., R.H. Gray, T.C. Quinn, N.K. Sewankambo, F. Wabwire–Mangen, D. Serwadda, L. Paxton, 1995. “Design and Feasibility of Population–based Mass STD Treatment, Rural Rakai District, Uganda.” Paper presented at the 1995 Annual Meeting of the International Society for STD Research, New Orleans, LA, August 1995.

Wishik, S.M., and K.H. Chen. 1973. *Couple–Year of Protection: A Measure of Family Planning Program Output*. International Institute for the Study of Human Reproduction: New York.

Woodhouse, G. 1995. *A Guide to MLn for New Users*. London: University of London, Institute of Education.

World Health Organization. 1989. “Contribution to the Working Paper of UNESCO.” Compiled for the World Youth Congress, Barcelona.

World Health Organization. 1993. *Indicators to Monitor Maternal Health Goals. Report of a Technical Working Group*, Geneva, Nov. 8–12, 1993.

Appendix A

Multilevel Model Regression Formats

Basic Cross–Sectional Model:

The basic cross–sectional multilevel regression model may be represented as follows:

$$Y_{ij} = \alpha + \beta P_i + \Gamma Z_i + \gamma X_{ij} + \delta P_i X_{ij} + \zeta Z_i X_{ij} + \mu_i + \varepsilon_{ij}$$

where:

Y_{ij} = the outcome variable of interest measured at the individual level (i.e., for individual j in community i) — for example, probability of conception in last 3 years, current contraceptive use;

P_i = variable or variables measuring program strength for community i;

Z_i = other community–level determinants of the outcome under study;

X_{ij} = individual– or household–level determinants;

μ_i = unobserved community–level factors (also referred to as “unobserved heterogeneity”)

ε_{ij} = unobserved individual–level factors; and

α, β, Γ, γ, δ, and ζ = parameters to be estimated.

For program evaluation purposes, the key regression parameters are the coefficients for the program variables (β and δ). The former coefficient provides a measure of the magnitude of direct effect(s) of program variables, while the latter provide(s) a measure of the relative importance of interactions between community–level program variables on the one hand and individual– and household–level variables on the other.

Multi–Equation Random Effects Model:

$P_i = \alpha + \beta Z_i + \lambda \mu_i + \varepsilon_{ij}$ (Program location equation)

$Yij = \alpha + \eta P_i + \Gamma Z_i + \gamma X_{ij} + \theta \mu_i + \varepsilon_{ij}$ (Outcome equation)

where:

P_i = variable or variables measuring program strength for community i;

Z_i = other community–level determinants of the outcome under study;

Y_{ij} = the outcome variable of interest measured at the individual level (i.e., for individual j in community i) — for example, probability of conception in last 3 years, current contraceptive use;

X_{ij} = individual– or household–level determinants;

μ_i = unobserved community–level factors (also referred to as “unobserved heterogeneity”)

ε_{ij} = unobserved individual–level factors; and

α, β, Γ, η, θ,γ, and λ = parameters to be estimated.

[Note: interaction terms have been omitted from the equations in order to simplify the presentation]

Fixed Effects Panel Model

The cross–sectional equations for two survey rounds may be written as follows (for the sake of simplicity, all interaction terms have been omitted):

$$Y_{ij1} = \alpha + \beta P_{i1} + \Gamma Z_{i1} + \gamma X_{ij1} + \mu_{i1} + \varepsilon_{ij1}, \text{ and}$$

$$Y_{ij2} = \alpha + \beta P_{i2} + \Gamma Z_{i2} + \gamma X_{ij2} + \mu_{i2} + \varepsilon_{ij2}$$

where:

Y_{ij} = time–varying outcomes;

P_i = time–varying program variables;

Z_i = time–varying community characteristics;

X_{ij} = time–varying individual characteristics;

μ_i = fixed unobservable community–level characteristics;

ε_{ij} = random error;

the “1’s” and the “2’s” refer to survey rounds; and

α, β, Γ, and γ are parameters to be estimated.

By differencing the two equations, we obtain:

$$Y_{ij2} - Y_{ij1} = \alpha + \theta(P_{i2} - P_{i1}) + \xi(Z_{i2} - Z_{i1}) + \phi(X_{ij2} - X_{ij1}) + (e_{ij2} - e_{ij1})$$

Because fixed parameters are invariant during the study period, they drop out of the difference equation. Of primary interest for program evaluation purposes is the “ϕ “ parameter, which measures the relative importance of changes in program variables in explaining observed changes in outcome variables during the time period studied. [56]

[56] Note that where the multilevel panel model is applied to successive surveys in the same sample clusters, changes in individual/household–level variables pertain to aggregate changes in these characteristics at the community level.