- PN-ABB-661

IMPROVING THE EFFICIENCY OF EDUCATIONAL SYSTEMS

THE EVALUATION OF EFFICIENCY IN EDUCATIONAL DEVELOPMENT ACTIVITIES

April 1986

IEES

Improving the Efficiency of Educational Systems

Florida State University
Howard University
Institute for International Research
State University of New York at Albany

United States Agency for International Development
Bureau for Science and Technology
Office of Education
Contract No. DPE-5283-C-00-4013-00

Improving the Efficiency of Educational Systems (IEES) is an initiative funded in 1984 by the Agency for International Development (AID), Bureau for Science and Technology, Office of Education. The principal goals of the IEES project are to help developing countries improve the performance of their educational systems and strengthen their capabilities for educational planning, management, and research. To achieve these goals, a consortium of U.S. institutions has been formed to work collaboratively with selected host governments and USAID Missions over ten years. The consortium consists of The Florida State University (prime contractor), Howard University, the Institute for International Research, and the State University of New York at Albany.

There are seven countries working with the IEES initiative to improve educational efficiency: Botswana, Haiti, Indonesia, Liberia, Nepal, Somalia, and Yemen Arab Republic.

Documents published by IEES are produced to promote improved educational practice, planning, and research within these countries. All publications generated by project activities are held in the IEES Educational Efficiency Clearinghouse at The Florida State University. Requests for project documents should be addressed to:

IEES
Educational Efficiency Clearinghouse
Learning Systems Institute
204 Dodd Hall
The Florida State University
Tallahassee, Florida 32306
USA
(904) 644-5442

Agency for International Development
Bureau for Science and Technology
Office of Education
Contract No. DPE-5823-C-00-4013-00
Project No. 936-5823

THE EVALUATION OF EFFICIENCY IN EDUCATIONAL DEVELOPMENTAL ACTIVITIES

DAVID W. CHAPMAN
DOUGLAS M. WINDHAM

April 1986

TABLE OF CONTENTS

List of Figures			v
Chapter One:	Introduction		1
	I.	Organization of the Monograph	1
	II.	Efficiency Criteria in Educational Assistance Programs	2
Chapter Two:	The Efficiency Concept		7
	I.	Definitional Issues	7
	II.	Operationalizing the Internal Efficiency Concept	10
	III.	Internal Efficiency in Education in the Developing World	13
	IV.	Constraints on Efficiency Enhancement	14
Chapter Three:	The Nature of Evaluation		29
	I.	The Meaning of Evaluation	29
	II.	The Timing of Evaluation	31
	III.	The Nature of the Evaluator	36
Chapter Four:	Criteria, Standards, and Indicators		39
	I.	Introduction	39
	II.	Confusion Among Criteria, Standards, and Indicators	42
	III.	Criteria for Evaluating International Technical Assistance Programs	45

Chapter Five: The Evaluation Process 57

I. Steps in Conducting an Evaluation 57

A. Formulation of a Point of View 57
B. Identification of the Purpose (Rationale) for the Evaluation 59
C. Identification of the Client, Sponsor, and Key Audiences 59
D. Identification of Resources and Constraints 61
E. Specification of the Evaluation Question 62
F. Formulation of an Evaluation Design 63
G. Selection of Data Collection Procedures and Collection of Data 65
H. Data Analysis 69
I. Interpretation and Reporting of Evaluation Results 70

Chapter Six: Technical Issues of Evaluation in the Development Context 79

I. Introduction 79

II. Issues in Level of Aggregation, Unit of Analysis, and Cross-Level Inference 79

III. Problems in the Measurement of Change 82

IV. The Limitations of Testing 85

V. Translation Procedures for the Cross-Cultural Use of Measuring Instruments 91

Chapter Seven: Context and Conclusions 97

I. The Political Context of Evaluation 97

II. Summary and Conclusions 99

Bibliography 103

LIST OF FIGURES

Figure One:	Simplified Model of Desired Program Impacts	55
Figure Two:	Eleven Steps in Conducting An Evaluation	58
Figure Three:	Suggested Outline for an Evaluation Report	73
Figure Four:	Examples of Back-Translation of Items from the Classroom Behavior Survey	93

CHAPTER ONE

INTRODUCTION

I. ORGANIZATION OF THE MONOGRAPH

This paper examines selected issues in the design and conduct of program and project evaluation in developing countries. While it focuses on the evaluation of international technical assistance programs that have the enhancement of educational efficiency as a goal, it has a much broader application--to programs sponsored by host government agencies, private sector enterprises, and private voluntary organizations. The objective of this paper is to improve the practice of evaluation in educational technical assistance programs where the issues of allocative and technical efficiency in the educational system are paramount.

This paper is intended for use by evaluators, program planners, administrators, and supervisory personnel in both host government and international donor agencies. It assumes that the reader has some professional background in program design, administration, or evaluation. The paper examines a series of practical issues in conceptualizing and designing educational efficiency evaluation studies that have particular relevance or pose particular dilemmas in international settings. It is not intended as a step-by-step manual for conducting evaluation nor is it a reiteration of basic issues in evaluation design or measurement. These subjects are widely available from other sources.

The monograph is divided into five separate but interrelated parts. In the remainder of Section 1 a brief introduction will be made of the evolving role of efficiency criteria in the evaluation of educational systems. Section 2 will provide detail on the nature of the efficiency concept as it is applied within educational programs (i.e., internal efficiency), its operationalization for the purpose of evaluation, and the special efficiency issues that exist in the developing world. This section will conclude with a discussion of the major constraints on educational efficiency enhancement efforts in schools in developing nations.

Section 3 discusses the nature of the evaluation process (as distinct from research). An emphasis is placed on the meaning of evaluation in terms of context and timing and the crucial role of the evaluator in determining the success of the evaluation process. Section 4 expands the discussion to examine the critical role of criteria, standards, and indicators in designing and conducting evaluation studies. A discussion of the special problems of developing and applying evaluation criteria within the international technical assistance system concludes this section. Section 5 then discusses basic procedural steps common to all evaluations. Section 6 focuses on the evaluation issues that are most troublesome in dealing with the topic of efficiency enhancement in

educational assistance programs in developing nations. The monograph concludes with a brief statement of summary interpretations and recommendations.

To accommodate a range of reader interest, the paper is presented at three levels:

1. The main text identifies and explores a series of evaluation issues and carries the primary argument of the paper.

2. Footnotes, referenced in the text, provide a more detailed discussion of selected points. The footnotes also provide specific references to other literature for readers who want to pursue more complete discussions of specific issues.

3. A bibliography of evaluation literature is provided at the end of the paper for users who wish to extend their reading in evaluation beyond the topics discussed here.

II. EFFICIENCY CRITERIA IN THE EDUCATIONAL ASSISTANCE PROCESS

Most non-military foreign aid to developing countries is awarded to fund specific development projects managed by the recipient government, by a donor agency, or by a development contractor. The question of how to allocate aid funds across competing sectors--health, agriculture, community development, education--is under continuous deliberation and debate. Educational development programs, as targets of international funding, have had a controversial record for several major reasons:

1. The outcomes of educational programs are long-term. The social and political pressures on both the host government and the donor agencies for quick evidence of program effectiveness have made educational initiatives controversial.

2. The rapid, short-term political, social, and economic changes that have characterized many developing countries often intervene to alter or cancel the effects that might have been achieved by educational projects.

3. Interventions in education are particularly susceptible to controversy because education is the most highly visible and participatory social institution and is involved with basic values of the society.

4. Education is a multi-input, multi-output system with an ill- defined definition of its process of production.

The increased interest in evaluation of international educational programs has resulted from the convergence of three factors:

Introduction

1. The last several decades have been marked by an increasing interest in assuring the responsible use of public monies. One expression of the increased public demand for accountability has been the mandating of evaluation in the legislation funding international development projects.

2. Social and educational programs go to the heart of the value system of a society and have complex impacts which go well beyond the obvious. Understanding the multiple impacts of such programs is important in project planning and administration.

3. As the level of financial investment in education has increased (under pressure from both market and social demand), the opportunity costs posed by educational expenditures in terms of other social investments foregone has increased. In most countries the growth rate in the education and training expenditure budget exceeds the average for government and often --with defense--education is the most rapidly expanding sector of government activity.

With this increased interest in evaluation has come a dramatic gain in the acceptance of a role for efficiency criteria and standards in educational evaluations. This gain in acceptance has not resulted from new conceptual insights by economists and finance specialists, but rather because of the budgetary crisis posed in many nations by reduced fiscal resources and a continuing expansion of social demand for education and training. As has been noted by Windham and Wang (1986), countries are faced with the choice of finding new fiscal resources (an economically unrealistic alternative), accepting quality deterioration and continued access inequity (a politically and ethically unacceptable solution), or increasing the efficiency with which educational resources are applied to the problems of instructional quality and access equity.

However, some economists, and almost all non-economists, have raised questions--if not outright objections--to the potential imposition of efficiency criteria and standards. Many of these critics see efficiency as a competing goal with the other, more qualitative, goals of education. This position indicates a basic misunderstanding of efficiency criteria and standards as tools for educational management and evaluation activities. It is a misunderstanding that has been encouraged by the activities of certain economists and financial analysts who have been satisfied with applying crude financial and other quantitative measures in research and evaluation at the exclusion of more central, but qualitative, measures. In addition, the quantitative standards that have been used often reflect a priority on ease of measurement rather than relevance to the political economy context of the educational process.

Stated simply, efficiency is not a goal of the instructional process; instructional goals must be defined in terms of intellectual, attitudinal, and behavioral criteria established by the appropriate polity--the student, the family, the school personnel, the community, voters, or government bureaucrats and policymakers. Until goals are

established, and, if possible, specified in some measurable way, efficiency has no meaning. Once generally acceptable goals are established, efficiency can be operationalized as a standard for resource allocation. The specific definitional issues related to educational efficiency will be dealt with briefly in the next section of this. For present purposes, it is sufficient that one recognize that efficiency is a criteria of instructional goal attainment, not an instructional goal in and of itself.

Understood in this context, efficiency is not something one need support or oppose; it is a natural resource allocation criterion for any public or private activity. What one can and should oppose is the often arbitrary and incomplete specification given to both instructional goals and efficiency criteria and standards. As the literature on educational efficiency shows, the ready acceptance of financial measures for inputs and of standardized achievement measures for outputs avoids the most critical aspects in the application of the efficiency concept, and its production metaphor, to instructional activities.

Education is a multi-input, multi-output activity. The latter characteristic, compounded by the subjective nature of many outputs, is the one that poses the greatest problems for evaluators and policy analysts. Even where one is willing to accept standardized achievement scores as an appropriate output measure, it should be obvious that issues of distribution are of as much social importance as the mean score or the proportion surpassing some acceptable minimum. In addition, education is believed to promote a variety of attitudinal and behavioral changes in students that may be complementary to certain achievement standards or may be acquired only at a sacrifice of the achievement standard. Where such outputs represent, at the margin, mutually exclusive goals of education, one must be prepared to make judgments of optimization, i.e., to order, in terms of preference, the various combinations of outputs from schooling. This is a sufficiently difficult process when one considers only two potentially competing goals (e.g., mean achievement and distributional equity), and it becomes increasingly complex as additional goals and multiple goal dimensions are considered. Since education in most countries is predominantly a public sector rather than a private market activity, the resolution of the goal establishment process may result in bureaucratic fiats and vague political espousals that can cloud rather than clarify the nature of educational goals. As a result, each district, school, or even teacher may have to make their own decisions about what goals (including their own job security and work environment conditions) they wish to optimize.

Two questions must be resolved by the education evaluator. The first is whose judgment on goal definition is to be accepted and the second is what model (or metaphor) of the education process will be used. If the efficiency criteria and standards are controversial, it is not because anyone is in favor of wasting educational resources. Rather, it is because individuals can have legitimate disagreements about the nature of the instructional process, specification of the goal matrix, and the value ordering of goal combinations.

However, the criticism that the efficiency criteria and standards ignore political "realities" is not acceptable. The last twenty years of educational expansion throughout the developed and developing world show that the failure of planners and policymakers to concern themselves with efficiency issues has created a harsh economic situation that defines the current political and social reality. Obviously, one must work within existing social and political structures and values in order to promote change. However, ignoring economic forces can only suppress and postpone their effect. Normally, the result is that the eventual economic crisis is even more politically and socially disruptive. Individuals who promote political or social realities over economic concerns are, in fact, promoting short-term realities over longer-term realities.

The remainder of this paper will shift the focus of discussion of efficiency to the evaluation procedures required to utilize the concept and to the special conditions of evaluating educational efficiency in developing nations. A major assertion of this paper is that the efficiency concept has generic relevance and that the basic methods and techniques of educational evaluation do not differ by level of national development. What does differ (and these are exceedingly important issues) are the problems of identification, definition, specification, and measurement of the appropriate variables to include in the education algorithm. It is not possible to resolve these issues generically because they are specific to a given political, social, and even cultural context. However, the discussion presented here is designed to clarify what can and cannot be done in terms of efficiency evaluation and, more importantly, to increase awareness of the political prerequisites for application of efficiency criteria and standards to education.

CHAPTER TWO

THE EFFICIENCY CONCEPT*

I. DEFINITIONAL ISSUES

Efficiency in education is often confused with two related but far from identical concepts: school quality and school effectiveness. School quality is probably one of the most diffuse and confusing terms introduced into policy discussions in the last twenty-five years. Depending on the writer, school quality can refer to input measures (aggregate expenditure, per-student expenditure, teacher qualifications, availability of facilities, equipment, and materials), process measures (teacher-student interaction, student time-on-task, peer effects, use of facilities, equipment, or materials), and output measures (test scores, promotion/graduation rates), or outcome measures (eventual social or economic success). Educational effectiveness, in contrast, normally is limited to output measures alone. Internal efficiency is the only concept, however, that links inputs to outputs in a systematic fashion. It is possible to have school quality and school effectiveness without having efficient operation of the school. The internal efficiency analysis asks whether more outputs could be achieved given the available inputs or, alternatively, whether fewer inputs could be used in providing the same level and mix of outputs. Thus, the internal efficiency concept is much more inclusive than those of quality or effectiveness and places a strong emphasis on the scarcity of resources and their appropriate utilization in schooling.

It also should be noted that the internal efficiency concept can be adapted to the inclusion of equity and access considerations. The equity and access measures (for example, participation and attainment by sex, size of place, region, or ethnic/racial group) can be included as output measures along with the more common achievement and attainment measures. Such an output definition is especially appropriate at a time when much of the policy debate deals more with the issues of aggregate access and less with the concern for social inclusion of underrepresented populations.

The efficiency concept and its role in the evaluation of education is best understood within the larger context of economic optimization. All optimization processes involve the maximization of the value of a given phenomenon (either a single item or a set of items) within the existing constraints of the environment. The maximization of profits, the optimization of social utility, and the minimization of costs are all examples of the generic optimization process.

---

* This section is based upon earlier work by Windham, "Internal Efficiency and the African School," prepared for the Institute for Research in the Economics of Education, University of Dijon, under World Bank Support.

Chapter 2

Economic efficiency is related but not identical to the more commonly understood concept of technical efficiency (as used in the study of physics or mechanics). Both efficiency concepts involve a ratio of an output or outputs to a set of inputs. In the case of technical efficiency, the ratio is stated purely in terms of physical quantities. Technical efficiency is optimized when one achieves the greatest possible ratio of outputs per unit of inputs. The procedure for dealing with multiple outputs and multiple inputs is conceptually the same even though the mathematics of the solution are substantially more complicated.

Economic efficiency includes all of the issues related to technical efficiency and adds consideration of the value of the inputs and outputs. This addition of values is required for decisionmaking in that the same physical quantities of different inputs may have dramatically different costs, and the same physical quantities of outputs may be valued quite differently by those who receive the outputs. If the technical relationships among inputs and outputs are known, the calculation of the most economically efficient combination of inputs can be derived if one knows the appropriate values (prices) to attach to the inputs and outputs.

The major problems faced in applying the economic efficiency concept--in practice--are the lack of or disagreement over values for inputs or outputs and the failure to consider alternative technological approaches. The first problem includes issues related to the propriety of using market prices for valuation, the difficulty of combining individual values into a group valuation or preference, and the inability to deal with purely subjective (psychic) benefits and costs. The second problem is one that, unlike the first, has not received a great deal of discussion by non-economists.

In theory, the process for determining economic efficiency involves three sets of decisions: the mix of outputs, the mix of inputs, and the technology to be used in transforming inputs into outputs (Bridge, Judd, and Moock, 1979). In a case where there is a single output and a given technology, the process of specifying the most economically efficient mix of inputs is quite easy if input values are given. However, unlike the manufacturing or private service sector, the social service sector (including government services such as education) rarely involves choices where outputs are singular or where the appropriate technology is obvious. Thus, the application of the economic efficiency concept in the social services sector has had to undergo several transformations.

At present there are four basic economic efficiency approaches that are used in public sector decisionmaking:

- Benefit-cost analysis;
- Cost-effectiveness analysis;
- Cost-utility analysis; and
- Least-cost analysis.

Benefit-cost analysis assumes that both outputs (benefits) and inputs (costs) can be stated in monetary terms. Since a common numeraire is used, the calculation of alternative benefit-cost ratios for different technological alternatives is possible. The technological alternative with the largest ratio of benefits to costs is considered the most efficient. Where benefits and/or costs are incurred over more than one time period, the present value of benefit-cost or rate of return approaches may be applied.

Cost-effectiveness analysis is used wherever it is possible to state input but not output values in monetary terms. However, cost-effectiveness analysis still requires that outputs be stated in quantitative terms. In education, such output measures can include test results, retention/dropout rates, attainment levels, numbers or proportion of students employed after graduation.

Cost-utility analysis relaxes the quantification requirements even further. While costs are still calculated in monetary terms, outputs are valued only in the subjective judgement of the decisionmakers. In the case of education, the decisionmaker may be a politician, a bureaucrat, an administrator, a teacher, a student, a parent or parents, or any combination of individuals to whom decisionmaking responsibility has been given. When one moves cost-utility analysis from the case of the single decisionmaker to that of group decisionmaking the subjective valuation is determined by the voting rules of the group. Even in the case of the single decisionmaker, the subjective valuation of output does not require explicit statement of the relative values of individual outputs in a multi-output situation.

Least-cost analysis involves the lowest level of conceptual sophistication of the analytical alternatives for measuring economic efficiency. It assumes that the desired outputs are given and requires only that evidence be presented that the proposed means of producing the outputs are the least costly of all feasible alternatives. Actually a subcategory of cost-utility analysis, the least-cost approach is used primarily in the determination of project design feasibility when there is a consensus that the benefits of the project justify its existence.

All of the economic efficiency approaches discussed here can be applied to either internal efficiency of schools (how well the schools achieve their stated goals) or external efficiency (how well the outputs of education match with social needs). The subsequent focus of this discussion is on internal efficiency issues. This emphasis is taken in full recognition of the reduced relevance of internal efficiency issues when the educational system fails to meet the external needs of society. Within the limits of the internal efficiency discussion, however, it will be possible to indicate several continuing controversies that exist in application of the efficiency concept to schools and school systems.

II. OPERATIONALIZING THE INTERNAL EFFICIENCY CONCEPT

The concept of internal efficiency as applied to education depends upon the input-output paradigm: efficiency exists where the value of educational output is maximized for a given cost of inputs (or where input cost is minimized for a specified value of output). However, attempts have been made in recent years (Thomas, 1977; Kemmerer, 1980; and Monk, 1984) to expand the analysis of efficient educational production beyond this simple model to one that includes the intervening technology by which inputs are transformed into outputs. In the following discussion, measures of educational quality will be reviewed at all three stages of the educational production process: input, technology, and output. However, as was noted earlier the concept of internal efficiency requires a linkage between the output and input measures.

The most common input measure used in the analysis of education has been per-student expenditure (or "unit cost" as it is commonly defined). This measure suffers from several methodological limitations (e.g., the assumption that all funds are expended in the most efficient manner) as well as measurement problems in the underlying definitions of expenditure or enrollment. The best uses for this measure are to compare among nations the willingness to pay for education or to pay for various levels and forms of education.

For example, Mingat and Psacharopoulos (1985) found that the per capita expenditures on education relative to per capita national product are 2.5 times greater at the primary level and 4.5 times greater at the secondary level in Africa than in other developing nations. Within Francophone sub-Saharan Africa, it was found that the per capita expenditure levels for primary education were only 20% of those in secondary education and 4% of those in higher education. The comparable figures for Anglophone nations were 36% and 2%.

Alternative input measures of educational quality include teacher quality (variously defined); availability and quality of facilities, materials, and equipment; and simple utilization ratios such as students per teacher, per class, or per school. There is inadequate space here to deal with the methodological and measurement limits on each of these concepts. These issues have been dealt with expansively in the past (Hanushek, 1977; Simmons, 1975) and more recently have been treated in the special context of developing nations (Fuller, 1985; and Psacharopoulos and Woodhall, 1985). It is sufficient here to note that at best these measures deal with the potential availability of instructional resources and at worst have little if any connection to the complex interaction of resources that takes place in the classroom.

The second set of quality variables--process measures--are intended to remedy part of the weakness inherent in input-only measures of quality. Examples of process variables are teacher-student interaction, peer influences, student time-on-task, curricular allocation, and measures of actual utilization of facilities, equipment, and instructional materials. The study of process phenomena is always more costly than that of the

quantified, input measures. Because process analysis is normally dependent on small sample sizes, it suffers from questions of generalizability of the findings.

While the process variables may well be better proxy measures for school quality than are the input variables, in isolation the process variables reveal little about either the costs or ultimate products of schooling. If education were solely a consumption process then one could justify slighting the outputs since, by definition, they would be largely identical to the process phenomena in a service activity such as education. However, to the extent that both individuals and societies view education as primarily an investment activity, continued attention is due to the output side of the school production function.

Many analysts view output measures as the only real measures of school quality; these individuals view all input and process measures as proxies for the actual output of the school or system. However, agreement on the importance of output measures has not been translated into agreement as to which output measure or set of measures is most appropriate. The output measures include cognitive, affective, attitudinal, and behavioral dimensions that range from mutually exclusive to jointly produced phenomena.

Such measures as examination achievement, attainment, graduation, and eventual social and economic success have all been used, singularly and in combination, as measures of school quality or effectiveness. In addition, it was noted earlier that access and equity measures could and should be included in any expanded form of the definition of school or school system outputs.

The weakness of the output variables is that while they can measure effectiveness (to what degree a previously stipulated goal is achieved), they are inadequate by themselves to allow for a judgment of efficiency. Many educational interventions (textbook distribution, modularized instruction, radio or television school broadcasts, or computer-based learning) may be effective at improving test scores or some other output measure but still not be efficient in terms of resource utilization (Levin and Woo, 1981; Kemmerer and Friend, 1985; Windham, 1983).

One of the few measures of school quality or effectiveness that has an efficiency dimension is cycle cost (expenditure per graduate of a level or "cycle" of schooling). While the unit cost concept measures only the level of expenditure per student (a measure of efficiency only if having children be students is a goal in and of itself), the cycle cost accepts the idea that an obvious function of education is to prepare graduates. A variety of formulae exist for calculating cycle costs dependent on data availability and quality (Dominiguez-Urosa, 1980). In most cases, the assumption is made that students who fail to graduate represent only a cost to the system and do not produce any personal or social benefits. More sophisticated models exist to deal with differential valuation of students by levels of attainment.

Another set of efficiency measures are those relating to school attrition and repetition (wastage). These measures assume explicitly that attrition or repetition is a negative aspect of schooling and attempt to use attrition and repetition rates as indicators of inefficiency. The problem is that both phenomena may be appropriate and necessary aspects of an efficient school or school system. For example, as an educational system expands access to poor or rural populations, attrition rates are likely to increase even if the input and process quality of school services is maintained. A school system may well decide to maximize initial entry into schools as a means of ensuring a minimal educational opportunity for all children. Attrition rates can only be used as a measure of efficiency if one is informed as to the access and equity goals of the system and the nature of the process by which the attrition decision is made. In addition, at higher levels of schooling, selection (forced attrition) may be an explicit process whereby limited school resources are matched with the most appropriate students.

Similarly, repetition within a grade level may be an appropriate instructional strategy for students who have fallen behind their initial cohorts. Unfortunately, the nature of attrition and repetition in most developing nations' schools can rarely be justified as part of a sophisticated educational strategy. Rather, they are the effects of instruction and examination systems that show little flexibility or adaptability to individual student needs. Thus, while they may be used as measures of school or system inefficiency, one is left to determine on a case by case basis the aspect of the school or system that is the source of the inefficiency.

The weakest linkage between quality and efficiency occurs at the process or technology stage. Normally, efficiency is defined for a given technology or, in rare cases, for choices among technologies given certain inputs and desired outputs. Only in the last decade have students of efficiency analysis carried their work to the classroom level. Although the methodological problems and time demands are extensive, this new direction offers the greatest opportunity for identifying potential means for increasing internal school efficiency in the long run.

In any deductive process in policy analysis, seven distinct steps must take place:

1. identification of conceptual determinants and effects;

2. definition and specification of the concepts;

3. operationalization of the concepts;

4. measurement of the operationable variables;

5. analysis of relationships among the variables;

6. interpretation of these relationships; and

application (or generalization) of these interpretations to general or other specific cases where the same conceptual determinants and effects exist.

The analysis of internal economic efficiency has proceeded with success only to the point of specification of the concept. As the above discussion indicates, there is great controversy over the operationalization of the internal efficiency concept and even wider disagreement over the standards for measurement, analysis, interpretation, and generalization.

In the context of developing nations these problems of methodology are magnified by the limitations on availability and quality of data. Budget definitions vary in nature and inclusiveness; school staff and enrollment data vary by country in terms of such factors as the sophistication of the data collection process and the time of the school year in which the enumeration of staff and student enrollments takes place. Because of these constraints, much of the current analysis of educational efficiency in developing nations must be qualitatively inferential. However, wherever possible, an attempt will be made to at least present examples of the quantitative dimensions of some of the problems that are faced by schools and school systems in attempts to create efficient instructional systems.

III. INTERNAL EFFICIENCY IN EDUCATION IN THE DEVELOPING WORLD

Assuming that a consensus were attained in regard to the issue of measuring internal efficiency, one would face a second major barrier to the analysis of the efficiency of schools in developing nations: the enormous variation that exists within and among the national "systems" of schooling. To even describe some of the sets of schools as part of a school system overstates the degree of coordination and supervision that exists within certain nations. In every nation, substantial divergence in the school environment exists along the dimensions of urban versus rural and developed versus underdeveloped regions. In some nations further divergence exists in terms of public versus private education, male versus female education, and secular versus religious education. When one adds to this complexity the ad hoc variations caused by cultural traditions, school administrator and teacher assignments, and a host of other factors, it is easy to understand the need for care in describing any example as an average or typical one.

The variation in educational conditions are dramatic within individual nations as well. For example, in 1981-1982, variation in primary school enrollment in Somalia, by region, had the following ranges for selected average regional characteristics:

Number of schools:	13 to 133
Number of classes:	39 to 816
Classes per school:	2.9 to 14.8
Total Enrollment:	1,064 to 38,719
Percent Female Enrollment:	20.8% to 48.5%
Percent Female Teachers:	6.4% to 56.9%
Student/Teacher Ratio:	19.7 to 63.0
Average Class Size:	18.4 to 47.6
Average School Enrollment:	53 to 704

Similar patterns could be found in almost any country selected. The point is that enormous variation exists in the reality of the classroom experience faced by students. To abstract from this reality is a necessary and appropriate device to allow for the analysis of those factors which are more common among schools and school systems. However, the analyst must never lose touch with the reality of the school environment in moving from the analytical to prescriptive aspects of policy analysis.

This paper makes an explicit assumption that a priority goal of developing nations is to improve the operation of their existing educational systems. While the merits and weaknesses of "deschooling" models will not be reviewed here, it is sufficient to note that the political and cultural acceptance of the traditional teacher-centered school model is not such that the model could be abandoned by most nations even if they had sufficient financial resources to do so.

IV. CONSTRAINTS ON EFFICIENCY ENHANCEMENT

The following section will discuss nine areas of constraint on the ability of developing nations to improve efficiency (both internal and external) in their primary and secondary schools. The purpose here is to introduce realism into the discussion of opportunities and strategies for efficiency enhancement. Too much program and project planning, implementation, and evaluation work has been conducted in the past without proper attention to these constraints. The result has been that hundreds of millions of dollars (both domestic and donor funds) have been expended in the last two decades without accomplishing the joint goals of expanded and more equitable access and the efficient production of the manpower needed for social and economic development. This emphasis on constraints is not a means to discourage further educational investments; rather, it is an attempt to increase the probability of successful intervention in the existing educational systems.

1. Political and Cultural Constraints. Educational systems may be the most conservative social enterprises that exist in developing nations. For all of the rhetoric from the educational extremists of various types, the individual school setting in developing nations is much the same as was the case at independence

and much the same as would have been found in Western Europe or the United States in the late 1800s.

In part, this may be due to the residual colonialist influence, but a more important determinant of the survival of teacher-centered, grade level instruction is the fact that the spread of credentialism outpaced educational development throughout the world in the 1940s and 1950s. As a result, politicians in developing nations have faced strong resistance from both teachers and parents in any attempt to move education away from the traditional forms of instruction and evaluation. As to the latter, the development of national or multi-national examination systems may have liberated nations from direct dependence on colonial testing systems but still advanced the institutionalization of the credentialing process.

While individual nations often have insisted on the need for a unique and locally-oriented curriculum for their schools, the need for internal and external standardization has restricted dramatically the ability to innovate in the educational system. Even those nations that engaged in more dramatic experimentation with curriculum have drifted back to more traditional educational systems.

In addition to this pattern of conservatism relative to dramatic reform, each developing nation has faced its own internal political and cultural limitations on the enhancement of educational efficiency. The roles of tribal, ethnic, and religious beliefs in the development of attitudes toward education often have been slighted in educational planning exercises. This indicates a need for the application of social marketing concepts to the attempts to remove social inequities in access to and retention in education.

2. Manpower Constraints. Given the political and cultural limitations on educational reform, the single most dramatic constraint on efficiency enhancement is the manpower situation. In the mid-1980s most developing nations, and especially those in Africa, are still at the beginning of their manpower development activities. Highly qualified manpower remains scarce even where the supply of highly certified manpower is increasing rapidly.

The manpower constraint has an impact on school reform in two basic ways. First, it limits the quantity and quality of individuals available to serve as teachers and, second, it determines the overall management efficiency of the society (including supervision and administration of education). In most developing nations from 25% to 50% of the primary school teaching force may be unqualified or underqualified. The lack of qualifications may refer to inadequate formal education (some primary school teachers are only primary

school graduates themselves), to a lack of pedagogical training, or to deficiencies in both areas.

As a result, the average primary school teacher may not be prepared to deal with school responsibilities except in a routine and repetitive manner. The infrequency, brevity, and frequent irrelevance of much inservice teacher training has limited this policy alternative in reducing the instructional impact of poor teacher quality. Of course, each country has a number of excellent administrators and teachers in primary education and there is less of a personnel quality problem in most secondary education programs. However, it also is true that those schools in rural and poor areas that require the most capable teachers consistently receive the least capable ones.

The issue of teacher salary and assignment policies will be returned to below in the discussion of incentive constraints. It is adequate to note here that there is little in the assignment, pay, and promotion policies of most educational systems to attract highly qualified individuals or to retain and motivate them if they are recruited. (In this regard the developing nation experience is different only in degree from experiences in the developed world.)

The second manpower constraint relates to management capacity. A shortage of individuals with research, analysis, administration, and supervision skills means that the individual school administrators and teachers receive little effective support from the central or regional offices of the education ministry. As a result, increased responsibility for the day-to-day operation of the school is often delegated to individuals unprepared to assume this responsibility.

A special manpower problem exists in the areas of science and mathematics. Even qualified primary teachers often have serious inadequacies in these subject areas. At the secondary education level developing nations face the same problems as those encountered in the developed nations, i.e., the opportunity costs are so high for anyone qualified in science or mathematics that they rarely become a teacher or, if they do, rarely remain beyond initial periods of bonding for loans or government subsidies.

3. Instructional Materials Constraints. The discussion of schooling as a teacher-centered process often fails to note that for a substantial proportion of the schools in developing nations the teacher is not only the primary but the sole source of instruction in the classroom. Much of the early literature on instructional materials dealt with the problem of localization of materials and the elimination of European or American ethnocentric biases. Unfortunately, a majority of schools in some nations would be willing to accept such materials if

they could be obtained, because at present they are operating without any materials.

In Liberia in 1983, it was found that a majority of classrooms had few if any textbooks and that nationally there was only one textbook for every twenty primary school students. In Somalia in 1984, it was determined that a shortage of 2,280,000 textbooks existed relative to what the national curriculum required for primary education. Even in Botswana, a relatively prosperous nation with a geographically concentrated population and good transportation infrastructure, a survey of schools in 1984 revealed shortages of textbooks and delayed delivery of instructional materials as consistent problems for primary education.

There are three distinct problems related to instructional materials for schools: development, delivery, and utilization. The development of instructional materials (including textbooks and instructional support supplies) ideally should be founded on the national curriculum for primary and secondary education. While large amounts of resources have been devoted by donor agencies to curricular reform and design efforts, many nations still operate with little more than a set of generalized objectives and vague goals. Issues of detailed content and sequence, the information most needed by the classroom instructor, are rarely available from the existing curriculum. Even where such detailed curricula exist, they often are not widely distributed to the teaching force.

With or without a curricular foundation, instructional materials development is further hindered by the scarcity of experienced indigenous authors and by the lack of a manufacturing capacity to reproduce sufficient quantities for national dissemination. The result is a continued dependence on foreign sources of supply or a prolonged period of materials development activity and an inevitable delay in materials being made available to the classroom.

As serious as the materials development problem may be, it is often overshadowed by the problem of distribution. One reason for the consistent inequity between urban and rural populations in educational achievement is that educational materials often are not distributed to the more distant schools. There are geographical, infrastructural, management, and manpower explanations for the distribution problem. Whatever the explanation, a failure to distribute available instructional materials is a source of major systemic inefficiency at the same time that it aggravates the problem of rural/urban and regional inequities.

Finally, in those fortuitous circumstances where instructional materials actually are made available in the classroom, the problem of utilization remains. Too often the distribution process represents little more than a "materials drop"

with teachers acquiring textbooks and instructional supplies but no advance instruction in their use. Without proper training or programmed instruction in the use of the materials provided, the effect of materials supply in the classroom will be minimized. Problems range from teachers who are uncertain about whether or how to distribute the materials to teachers who decide it is simpler to continue their teaching as before and ignore the new materials. Any program of intervention based on the current literature's confidence concerning the efficacy of instructional materials (Heyneman, 1982) must take into account the three aspects of development, distribution, and utilization.

4. Facilities Constraints. The condition of education in developing nations often can be startling. For example, Heyneman (1983) found that in Malawi in 1979 only one pupil in eight had a chair and only one in eighty-eight had a desk. He notes:

> ...walls frequently collapsed after a rain; roofs had large holes; wind and storms disrupted class activity as a matter of course. The normal classroom was dark and stuffy; students sat on the ground, balancing an exercise book or slate on their knees.

A similar environment for students may be found in many parts of the developing world and for some nations, this environment represents the modal learning environment in rural areas.

A recent World Bank survey of research (Fuller, 1985) confirmed earlier analyses that examine the role of facilities quality in determining student achievement. While correlations are found between school building quality or availability of special use facilities (libraries and laboratories) and student achievement, these correlations tend to be small and of questionable significance. While some minimum facility quality undoubtedly is required in most environments, and there is a persuasive case to be made for facilities quality as a constraint on school learning, there is no similar case to be made, intuitively or statistically, for facilities construction as a major vehicle for efficiency enhancement.

The status of facilities utilization is a more critical issue than the simple availability of schools built to Western standards. The availability problem can be dealt with in the short run by adaptation of facilities designed for other purposes. The 1978 National Education Survey in Liberia found that 43% of the schools were operating in facilities originally designed for other purposes. A significant number of schools already in operation in other countries are sited in facilities that meet minimal if not optimal structural requirements. Even in a

case such as Botswana, where 27% of primary school classes are held outside a formal classroom building, this is not as serious a constraint on learning as it may appear. Given the choice of receiving instruction in an overcrowded, poorly designed building or in the open air, many teachers and students will choose the latter where climate and custom permit this alternative.

The question is not whether there is a shortage of facilities given Western standards. Obviously such a shortage does exist in both urban and rural environments. The relevant efficiency issue is whether construction of an improved facility will enhance learning. Given that the ceteris paribus conditions often include unqualified teachers, little, if any, instructional materials, and no clearly disseminated curricular format, the skepticism toward facilities development as a solution to the inefficiency and poor quality of education appears justified.

The irony here is that facilities development has been the major single focus of bilateral and multilateral assistance to education in developing nations over the last quarter-century. This assistance has aided and encouraged the quantitative expansion of schooling at the same time that significantly less attention has been directed to the internal classroom operations of either the existing or new schools. Only if one accepts a singular goal of providing wider access to poor quality education can these narrowly-based facility development projects be countenanced as an appropriate assistance strategy.

In recent years facilities programs have responded to some of the common criticisms of these endeavors. Many of the examples of new school construction incorporate low-cost designs, use of local materials, and a low-maintenance requirement. Even at their best, however, facilities programs create a preferable precondition to efficiency enhancement but do not qualify as a sufficient (and perhaps not even necessary) precondition.

The long-term solution to the facilities problem is going to require a mobilization of local rather than national or international resources. Such a policy shift will involve loosening or abandonment of national construction standards and the possibility for continued differences or even inequities in facilities quality among regions or individual schools. However, a locally-oriented responsibility for school construction and maintenance would promote efficiency by increasing the number of schools that meet at least the minimum standards required of facilities. In addition, such a reorientation of responsibility would free other funds to be used for more direct means of enhancing quality and efficiency.

5. Incentive Constraints. The systems of primary and secondary education in developing nations exist within sets of complex configurations of incentives. These incentive sets range from the employment and wage or salary incentives of the national labor market to the specific behavioral incentives that affect teacher and student performance in the classroom (Windham, 1980).

Two major problems exist relative to incentive effects in education. First, do planned incentives have their desired effect and, second, do unintended incentives exist that promote counter-productive behavior? An example of the first problem exists in terms of the paucity of information available to the student, family, and often even to the teacher on labor market requirements for school leavers. In many cases, even the requirements for advanced academic, vocational, or technical education are unavailable to students and families in making their choices of academic programs.

An example of unintended incentives exists in the bureaucratic system which requires similar or identical pay for teachers regardless of their subject specialization or job location. A failure to provide salary or other incentives to those teachers with skills marketable outside the teaching profession (science and mathematics specialists or vocational/technical craftsmen) will mean that the system will face a continuing shortage of such personnel. A failure to provide salary supplements or other compensation for teachers assigned to rural areas will lead to shortages and/or a rapid turnover of personnel assigned to such schools.

Educational planners need to be concerned with the incentives for individuals to become and remain teachers; in many cases the incentives are stronger to become a teacher trainee than to remain a teacher after the training period is complete. With training stipend, food, and housing provided while in training, some trainees may face a lower real income after graduation than before. Furthermore, when the transition to first teaching assignment is delayed because of the bureaucratic appointment and payment process (a transition that takes most of one year in some systems); many of the best teacher training graduates find uses for their skills outside the teaching profession. Even where bonding regulations and required periods of service are strictly enforced, one only postpones the inevitable attrition of the best personnel from the teaching force. When able teachers depart from the profession, a part of the training investment is wasted and the schools remain without qualified personnel.

One of the most disturbing effects of the incentive structure in education is the convergence of conditions that lead to discouragement of able and motivated

teachers. Some of the constraints noted above--facilities, instructional materials, and community attitudes--can impose a harsh burden on a new teacher.

The nature of teacher assignment policies is such that new teachers--who are in the most need of on-the-job support and guidance--frequently are assigned to the most difficult schools. Some new teachers find themselves in single-teacher, multi-grade schools in areas where culture, religion, and even diet may be dramatically different from their own. The results range from poor motivation to high absenteeism to outright abandonment of the school by the teacher.

The design of effective incentives for any education system is an evolutionary process (Green, 1980). It requires recurrent review, analysis, and reform. However, with the exception of changes in salary levels, little explicit attention appears to have been paid by planners and administrators to the incentive phenomena as sources of efficiency constraints.

6. Attitudinal Constraints. Schools and school systems in Africa face a special set of constraints in terms of the standards and expectations of administrators, teachers, parents, and students. Each actor in the school process may and probably does view the process in a different manner. The administrators are concerned primarily with issues of stability and quantitative standards of performance; the teachers are concerned primarily with the behavior and academic performance of those students within their direct responsibility; the parents are concerned with the achievement of their child in a relative as well as absolute sense; and the individual pupils present a vast array of personal concerns that are unlikely to be fully congruent with those of any of the other individuals

Birdsall and Cochrane (1982) hypothesized that family perspectives toward schooling were due to three sets of influences. These were household factors (parents' education and income), economic environment factors (school costs, wage rates, returns to schooling), and factors related to what were called "unobserved preferences." These preferences were assumed to be a function of social norms, family structures, and culture.

These preferences become the source of the variation in accepted standards of behavior and academic performance that occur even in a single community school but are a major factor in a national educational system. Part of education's traditional "hidden agenda" has been to bring a greater standardization to the range of attitudes that parents and children have toward schooling and other social processes.

An example of how these preferences and attitudinal factors can act as a constraint is the difficulty of introducing objective evaluation into a community with an explicit hierarchy of social status. The teachers who assign grades based on school performance alone may find themselves under strong pressure from the community elite. The uneasy acceptance of meritocratic bases for assignment of social roles that one finds in Western society is not always reflected within the village life of rural Africa or Asia.

Attitudinal factors also have a strong deterministic role in how well teachers accept proposed instructional innovations (Benyahia, 1983). There may be a strong resistance to experimental learning systems for primary education if the result is greatly increased time demands on teachers. A more dramatic attitudinal effect has been observed in the frequent resistance by inexperienced teachers to national dissemination of television or radio instructional programs.

The incidence of failure of these programs (in terms of dissemination if not experimentation) is due to an inability or unwillingness to appreciate the teachers' strong preference for control of their own classroom and the teachers' fear that the new technology will become a substitute rather than a complement for the traditional role of the classroom teacher. The new generation of instructional technology (involving calculators and computers) will face similar resistance if planners and implementors do not include consideration of attitudinal constraints in their strategy for efficiency enhancement.

7. Management Constraints. The manpower problem as a limitation on management capacity was discussed earlier. In this section, the focus is on the structural and bureaucratic factors that limit educational efficiency enhancement in developing nations. In addition to the shortage of trained manpower, the major managerial constraints on educational efficiency stem from: (a) an inappropriate information and incentive system; (b) the lack of explicit and quantifiable goals; and (c) the state of technological advancement in the area of

Most education ministries operate with a hierarchial decisionmaking system headed by a minister who is more likely a political official than an educational professional. Most procedural decisionmaking is concentrated at the level of the permanent secretary or director general; this person is normally the senior professional in the system. The nature of information and incentives in the developing nations is such that an excessive amount of decisionmaking is placed at the level of the permanent secretary. Among the reasons for this are the inadequate training and experience of subordinates, the reluctance by subordinates to bear responsibility for decisionmaking, and the desire by senior officials to control even routine ministerial operations. The result of this process is that

delays occur; the ultimate decisionmaker is further removed from the actual event, and thus, often less well informed than a subordinate decisionmaker would be; and no one is left with time available to deal with the long-term planning concerns that should be the primary responsibility of the senior administrator (Windham, 1982). The problem in most developing nations is not that educational systems are hierarchial, but that there is not an efficient allocation of authority and responsibility among the levels of the system.

Any management system would suffer from a lack of explicit goals. Accountability requires that both the practitioner and administrator agree as to the desired outcomes of the system. Ministries of education serve many functions in addition to that of instruction: they are major sources of public service employment, they are the most widely disseminated examples of central government largesse, they may represent a political network of government representatives, they are distribution points for information and propaganda, they are day-care centers for children of the urban employed, and they are centers of community activities. With such a multi-output institution and with no indication of the rates of tradeoff among these outputs and the multiplicity of the specifically instructional outputs, the constraint on management evaluation is obvious. The result has been that easily quantifiable factors--number of schools, number of students, number of teachers, pass rates, attrition/repetition levels, and examination scores--have dominated in the formal evaluation of educational management.

The third facet of management constraint is the state of technological advancement. In most educational agencies, the quality of data collection (as rudimentary as it may be) is far superior to the analysis and dissemination of data. At a time when the availability of microprocessing equipment is increasingly affordable, many planning units continue to work with desk calculators or to wait for infrequent access to mainframe computers. The need for databased decisionmaking is an obvious one, but is restricted by the lack of accuracy and timeliness with which data analysis can be conducted. An additional need in this area is for more and better training of policy analysts in doing iterative provisional analysis of data in the time frames normally encountered in ministry work. Traditional conservative research techniques simply are not always applicable to the time frames allowed for much of the policy work done in government ministries and research support agencies.

Infrastructure Constraints. For someone who has not had the experience of field work in rural Africa or Asia, the constraint that is easiest to overlook is that of the infrastructure (i.e., the roads, highways, telephones, and other communication systems that are taken for granted in more developed societies).

The nature of the environment in certain sub-Saharan nations is such that a significant number of schools cannot be reached by road vehicles for several months each year because of the effects of seasonal rains and the consequent flooding. The geographic isolation of certain other schools makes them difficult to visit at any time during the year.

The condition of roads and highways (where they do exist) are normally such as to require much greater time and energy for travel than for the same distance in a developed nation. Telephone and other telecommunications systems are well developed in cities such as Jakarta and Nairobi, but elsewhere, even in Indonesia or Kenya, one will face uncertain availability and unreliable quality of service. In less wealthy nations, the telecommunications contacts can be unreliable even in the capital cities.

A special constraint on the use of the new informational technologies involving computers and related equipment is that machines have to be adapted to deal with both power surges and failures. The result is that the cost of installation, maintenance, and operation of such equipment is higher than in Western Europe or the United States. A more generic problem is the lack of a repair and parts replacement system for all types of equipment from vehicles to computers.

The purpose here is to stress the danger of false assumptions about what can be done in the implementation and administration of efficiency enhancement projects. All designs of reform efforts must be predicated on the probability of delays in delivery and communication. Any project involving interaction between central personnel and schools will have to take into account the serious infrastructural barriers that exist. Project designs in education have been consistently underfunded for both implementation and evaluation activities. The history of educational projects in developing nations is weighted with failures; however, a majority of those failures were caused as much by problems with the implementation design as with the behavioral conception of the projects. Such failures will continue as long as project conception and design is undertaken by individuals unfamiliar with the realities of the social environment and especially the constraints imposed by the nature of infrastructure in urban and rural areas.

9. Donor Assistance Constraints. To this point the discussion of constraints of efficiency enhancement have concentrated on the indigenous limits within developing nations to attempts at educational reform. It is only fair to direct some attention to the external influences that have led to some of the barriers to efficiency one encounters. The effect of the colonial heritage was mentioned in passing in an earlier section; in many nations European systems of education

have been adopted. In some cases the adoption has been ad hoc and at other times complete with curricular standards and examination systems (Watson, 1982). The concern here is not with the oft-stated questions of the ethical propriety of this cultural intrusion but rather with its functional propriety.

Developing nations, often operating with per capita income levels comparable to those of the late 1800s in most of Europe and the United States, have been expected to mount educational systems nearly contemporary with those of the donor nations. In addition, programs of social inclusion for rural populations, women, ethnic and religious minorities, and the physically and mentally handicapped have been urged on these nations by representatives of societies that themselves have only begun to deal with these issues. One does not need to be a dedicated student of history to recognize that programs of social inclusion in the West followed rather than inspired the major periods of economic development. In fact, to the extent that the educational inequalities of the 1800s promoted large scale capital accumulation, there is a legitimate question as to whether the current levels and types of educational expenditure in developing nations do not represent potential restraints on, rather than sources of, economic development.

These countries are being asked to serve as an experiment to test whether development can occur without the concomitant inequalities that have existed elsewhere in the past. The nobility of this goal is slighted, however, by the fact that since the early 1950s little success has been achieved in either economic growth or social inclusion. As noted earlier, when it has occurred, the social inclusion success has often meant simply that wider access has been gained to a school experience of marginal if any value.

The concern here is less with the strategy of the national leaderships and more with the hypocrisy of the donor agencies. The developing world has served as an experimental laboratory for everything from modularized instruction to "lifelong learning." Long run incremental strategies for educational development have been sacrificed to allow for the ad hoc interventions of Western educators. The attention span of domestic politicians and donor administrators has been such that these experiments--many of which had potential for improving school instruction or system performance significantly--were rarely translated into fully disseminated systems. As a result, one educational novelty has succeeded another with little evidence of an accumulation of experience or wisdom. The facilities emphasis of donors is one of the few examples of a long-term orientation in donor policy. While even these activities have been idiosyncratic within individual nations, the attractiveness of facilities projects in terms of

finite obligations and visible signs of accomplishment has made them one of the rare long-term strategies evidenced by most donor organizations.

In addition to the factors mentioned above, the most common characteristic noted in regard to donor behavior is the lack of inter-donor coordination of activities. Although substantial progress has been made in regard to donor coordination in the last five years, the continuing fragmentation of donor efforts has had two major negative effects.

First, the development plan for education in a developing nation is less likely to be an intuitively-derived strategy on the part of the host nation's planners and more likely to be a catalog of those activities donors have expressed a willingness to support. Even where a systematic independent educational development plan is produced, the implementation of various parts of the plan soon becomes dependent upon the garnering of donor support. The need for matching funds for donor-assisted activities leaves little domestic capital for support of other development activities which have not found favor within the donor community.

The second negative aspect of donor fragmentation is the effect of uncoordinated program initiatives on recurrent cost obligations of the host governments. Even with grant contributions and concessionary loan terms, the host government often remains burdened by significantly higher cost obligations as a result of donor activities. Increased recurrent cost burdens are a dramatic characteristic of facilities expansion and teacher training initiatives. The latter can be especially problematic in that the host government is left with the cost of continuing the new preservice or inservice training programs while incurring new or increased salary obligations for the teaching force.

The effect of the constraints imposed by donor behavior can be offset in part by a greater exercise of discipline and authority on the part of the host governments. There needs to be a greater willingness to say "no" or, alternatively, for the host government to play a more active role in the design and justification of project activities. Also, increased coordination should not be understood to mean only coordination among the donors, but improved coordination with the host government's long-term educational plans. Otherwise donor cooperation may be viewed as a conspiracy among the donors against the interests of the host nation. The ultimate goal of any truly coordinated program should be to develop a full and equal partnership between the donor community and the host nation to replace the present advisor/client relationship that exists in a majority of developing nations.

10. Financial Constraints. The discussion of financial constraints has been left to last, in part, because they are the most obvious constraints. However, it is more important that it be understood that alleviation of the financial constraints will do little to improve the educational system unless the other aforementioned constraints are dealt with as well. The solution to educational problems is not likely to come--or should it come--simply from more funds being made available. The solution must be found in the more efficient use of the resources already invested in the system. Once efficiency in the use of resources is achieved, it will become easier to justify greater resource requests for education and the funds allocated will be assured of having a greater effect on school and school system outcomes. For the last two decades, new expenditures have been used to remove or disguise the effect of the school system's inefficient design and operation. In the next two decades efficiency enhancement should become a prerequisite for new allocations of funds.

The debate over new funding versus efficiency enhancement may be moot in most countries for the remainder of this century. The vast majority of the developing nations do not have the choice of using large, new allocations of funds for education. Aggregate economic stagnation combined with increasing demands from other social sectors (especially in the areas of health and population) and from the economic infrastructure will force most nations to choose between increased efficiency or a further qualitative (and perhaps even a proportional quantitative) decline in educational services.

The largest source of funds for education remains the host nation. The largest item of expenditure will remain teacher salary costs. The needs of the society are not served either by increasing the quantity of unqualified teachers or by simply raising the pay of the existing population of unqualfied teachers. Any analysis of fiscal investment alternatives or efficiency enhancement activities must begin with the reality of the teacher-centered classroom process. The challenge for the remainder of this century is to increase the efficiency of the teacher-centered process within the wide range of constraints discussed here.

CHAPTER THREE

THE NATURE OF EVALUATION

I. THE MEANING OF EVALUATION

Contemporary arguments about the meaning and role of evaluation consider evaluation to be a systematic activity undertaken to assist some audience to judge and improve the worth of a program or activity. This definition encompasses four key dimensions of evaluation.

1. Evaluation involves judgments of worth. Evaluation entails a value judgment about program worth; this is the goal and distinguishing characteristic of all evaluation. Evaluation may play different roles, such as formative or summative, but these roles have to do with the social uses of evaluation and have nothing to do with the nature of the evaluation process itself.

   Judgments are truth claims offered in the absence of decisive evidence. The adequacy of a judgment is determined by the sufficiency of the grounds for that judgment. These grounds consist of evidence, beliefs, and interpretations that people hold to be relevant to that evaluative judgment.* Debates over the adequacy of an evaluation are debates over the sufficiency and relevance of the factual and intuitive grounds used to support the claim of a program's worth.

2. Evaluation is different from research. Evaluation and research are both forms of systematic inquiry, sharing a number of techniques, methods, and procedures. They both play an important role in program development. They are, however, significantly different activities. The most essential distinction lies in the purpose the two activities serve.** These differences concern: (a) the focus of the inquiry, (b) the generalizability of results, and (c) the role of valuing.

This argument draws heavily from the recent work of Edward F. Kelly, "Getting Value in Evaluation," School of Education, Albany, N.Y.: State University of New York at Albany, 1985. See also E.F. Kelly, "Evaluation: Issues and Practices," School of Education, Albany, N.Y.: State University of New York at Albany, 1983.

** The distinction between evaluation and research is developed well by J. Popham, Educational Evaluation, Englewood Cliffs, N.J.: Prentice Hall, 1975. The distinction blurs somewhat in discussions of applied social research, policy analysis, and policy studies. However, a further elaboration is not necessary to convey the points under discussion in this section.

Research (in its positivistic as opposed to normative or prescriptive sense) is undertaken to produce new knowledge. It is guided by theory and investigates why things happen as they do. The purpose of research, therefore, is to provide generalizable findings that have applicability to other settings. Further, the purpose of research is not to impose value judgments, but to expose systematic relationships and patterns.

An explicit purpose of evaluation, on the other hand, is to yield judgments of the worth of a program that contribute to decisions about the program's design, administration, effectiveness, and efficiency. Evaluation is a practical activity which is guided by the questions of concern to the stakeholders of the program being evaluated. Consequently, generalizability, which is a cornerstone of research, is not always a primary goal in evaluation. Strong causal claims are not a necessary part of evaluation, nor are results that generalize to groups beyond the project at hand. Contributing to decisions about a specific program, however, is the key outcome desired of an evaluation.

3. Evaluation contributes to decisionmaking. Evaluation is performed in the service of decisionmaking. The purpose of evaluation--of determining the worth of a program or activity--is to provide useful information for decisionmakers to choose among policy alternatives. Decisionmakers include more than project planners and administrators; they include other groups affected by the existence or operation of a program.

The primary audiences for an evaluation are: (a) those with the greatest opportunity to make use of the results in modifying the program or the environment in which the program occurs and (b) those with the greatest need to confirm their own response to the program. The program sponsor, planners, and administrators are frequently the most salient audiences of an evaluation study. They generally are the ones who commission the evaluation study and are the recipients of the final evaluation reports.

However, other groups may have substantial interests in the program and may face decisions of their own: whether to continue participating in or lending political support to a program, whether to compete with the program for resources desired by both groups, and whether to adopt portions of the program in other settings. These groups also are "stakeholders" in the evaluation. They have a vested interest in the outcomes of the evaluation and are faced with their own sets of project related decisions.

Evaluation is a practical activity leading to action.* Evaluations are practical arguments that lead to action rather than new knowledge (Kelly, 1985). They are arguments in the sense that evaluations posit a series of premises that lead to the evaluative conclusions. The premises of the evaluative argument are composed, in part, of evidence, beliefs, and interpretations in an explicit value-laden context. The product of the practical arguments of evaluation is action, while the product of the theoretical arguments of research is hoped to be new knowledge. This is not to say that new knowledge cannot result from evaluations, but the generation of new knowledge is not the purpose of the activity.

II. THE TIMING OF EVALUATION

Evaluation can occur at different times during the planning and conduct of a program. When evaluation occurs relates closely to what role evaluation plays and the types of decisions to which the evaluation contributes.

1. Conducted in advance of a project, during or preceding the design stage, evaluation provides a mechanism for identifying issues, constraints, and points of potential program intervention. This role, in which evaluation operates as policy analysis, is frequently described as "needs assessment."

2. Conducted during a project, evaluation provides a means for building a self-correcting process into the program. In this role, evaluation provides information for the necessary readjustments in the operation and conduct of a program. It also provides warnings about policies, procedures, and program elements that may have unforeseen negative consequences before such consequences reach damaging proportions. This normally is called "formative evaluation."

3. Conducted at the end of a project, evaluation yields information for longer-term decisions about whether a program should be extended or terminated, maintained or revised, disseminated to other locations or abandoned. This process is termed "summative evaluation."

Donor-funded international programs normally mandate formal evaluation in a summative role to render a judgment of project success or failure at the conclusion of the project. To date, however, evaluation in international technical assistance programs

E.F. Kelly presents some of the clearest analysis of the evaluative argument and its relation to practice in a recent series of papers referenced in footnote on page 29.

has proven most useful when it has occurred early, in the role of policy analysis and needs assessment. Evaluation that is conducted after a program is completed, while still potentially useful, occurs too late to assist the present project. The usefulness of summative evaluation results in subsequent project design efforts is limited by: (a) the infrequent replication of large-scale technical assistance programs and (b) where replication does occur, the need to commit to continuation or to redesign of projects long before summative data about the earlier project cycle is available.

This poses a paradox in program planning. The clearest mandate for evaluation (a summative role) would have it occur after evaluation's point of maximum usefulness (a formative role). This will be discussed more fully later.

A useful framework for thinking about the roles evaluation can play in program planning and operation is offered by Stufflebeam (1971). He distinguishes among context, input, process, and product evaluation in terms of when the evaluation activities occur in the program and what types of evaluation questions they address.

Context evaluation is undertaken to identify the current conditions, issues, opportunities, and constraints in the environment. It is a form of needs assessment, an early activity to identify the types of programs appropriate to a given setting.

A context evaluation initially involves identifying the limits of the domain to be served. Data are collected to identify current conditions, unmet needs, and unused opportunities, as well as the problems that may limit response to those needs and opportunities. When possible, the analysis draws from existing sources of data; however, additional empirical studies also may be necessary to gather the needed information.

The information provided by the context evaluation contributes to several types of decisions: (a) the setting to be served, (b) the general goals to be sought, and (c) the objectives to be achieved. Context analysis serves as the background for more specific project design activities that may follow.

Context analysis in international programs is illustrated by the sector assessments as undertaken by the World Bank or USAID.* The purpose of these studies is to

World Bank sector analysis procedures are described in Baum, W.C. and S.M. Tolbert, Investing in Development: Lessons of World Bank Experience, Washington, D.C.: The World Bank, 1985. Sector assessment techniques used in recent studies sponsored by USAID are described in Cieutat, V.S., "Planning and Managing an Education Sector Assessment," Office of Science and Technology/Education, Washington, D.C.: United States Agency for International Development, 1983, and Cieutat, V.S., Planning and Managing an Education Section Assessment: Update, McLean, Va.: Institute for International Research, 1986. See also, Robinson, B. "On Methodology for Education Sector

examine and assess the resources, existing plans, needs, problems, and opportunities ir individual sectors of the economy. For example, the Education and Human Resource (EHR) Sector Assessment in the Yemen Arab Republic (Government of the Yemer Arab Republic, 1985) describes the EHR sector within the larger economy, discusses the current and projected plans and resources of the Yemen government, presents data on the current condition of EHR activities in Yemen (e.g., enrollments, student flow, staffing facilities, etc.), discusses opportunities for the continued development of that sector, and proposes specific recommendations for how continued development might proceed. The study preceded any particular project plan or investment strategy. It was undertaken as a background study of the context in which future EHR activities would occur.

But why would a government, with an education ministry in daily contact with the sector, need to undertake such an analysis? One reason is that the ministry's activities relate to the resolution of particular problems or the implementation of particular policies. Ministries conduct studies of pressing problems but seldom have the opportunity to establish an overview of the EHR sector as a whole.

Another example of context evaluation in international settings is development of the USAID Project Identification Document (PID) which provides a cursory needs assessment, a policy analysis, and a broad- based rationale for a project in a given program area. It is focused on a particular project rather than an entire sector, but it still precedes specific project design. Obviously, project identification activities are greatly advanced if preceded by a sector assessment.

Input evaluation provides information for determining how to utilize resources to achieve project objectives. It involves identifying and assessing: (a) relevant capabilities of the responsible agencies and groups, (b) strategies for achieving project objectives, and (c) designs for achieving specific strategies. The information provided in an input evaluation is essential for structuring specific designs to accomplish project objectives.

The degree to which input evaluation operates as a formal activity varies by sponsor and by project, though all local ministry and donor agency projects go through some form of planning which considers the issues addressed in input evaluation. An example of input evaluation in USAID is represented by the Project Paper stage of project development. During the Project Paper, the project concept, identified in the PID, is elaborated and a specific project design is proposed. This activity involves assessing the anticipated capabilities of relevant groups, appropriateness of alternative strategies, and feasibility of the design proposed. Inputs are evaluated in terms of least-cost, cost-utility, cost-effectiveness, or cost-benefit criteria.

Analysis," Washington, D.C.: United States Agency for International Development, 1973.

Process evaluation is synonymous with what is widely called formative evaluation. Its three main objectives are: (1) to identify unanticipated consequences of the projgram in time for program managers to avoid those that are undesirable; (2) to provide ongoing information on the performance of the program (e.g., the extent of its implementation, fidelity with initial goals, problems in the way the program is being fielded or received); and (3) to document what is happening in the project.

The importance of this last function is greater than it may seem. Programs and projects often are fielded with little careful documentation about implementation activities. How much of what type of material or service was received at what time by which participant? Too often programs are concluded to be successful (or unsuccessful) without describing the relevant dimensions of what occurred--such that later efforts to replicate (or avoid) the successful (or unsuccessful) components are impeded by ambiguity about what really occurred. Programs are poorly served by evaluations that render conclusions about the effectiveness and impact of the program, but fail to document the activities and events which in fact constituted the program.

The chief audiences of process, or formative, evaluation are those in positions to make the necessary mid-course adjustments that may be needed. This is usually the program manager or, in some cases, the program sponsor. Process evaluations tend to be presented in less formal ways than context, input, or product evaluations. The emphasis here is on continuous, timely information on what is currently happening. Consequently, reporting formats are more likely to involve informal memoranda and frequent conversations than formal written reports.

Curiously, while international technical assistance programs tend to have strong context and input evaluations, they tend to be weak in process evaluation. Several factors contribute to this tendency. First, altering initial commitments and contracts to solve mid-course problems often is difficult. Making such alterations requires considerable justification and extra work, and it raises questions at higher administrative levels about the adequacy of the initial planning process. Further, in well-designed programs, all the components are interrelated. Changing one factor may require a much wider set of changes in other factors. Formative evaluation results sometimes are ignored (or never sought) because of a tacit belief by program managers that following the original plan, even if flawed, is preferable to undertaking changes that either may reflect poorly on the initial plan or appear of questionable cost-effectiveness given established procedures and fixed costs.

Second, large-scale international projects operate in contexts bounded by strong and active interest groups. A change in a program's operation may offend some groups even if it would appeal to others. The lack of overt attention paid by project staff or ministry personnel to formative evaluation is sometimes linked to a concern for offending important interest groups.

Third, project design procedures usually call for special teams to conduct context, input, and product evaluations, but typically, they do not specifically provide for process evaluation. Rather, this is left to the project management team. In practice, it is usually assumed that monitoring activities of the sponsor and the regular administrative duties of the program staff will detect needed changes. Unfortunately, this is not always the case. Sponsors have their own political and bureaucratic agendas, and these may be concerned more with ensuring that implementation and fund commitments occur on schedule than with fine- tuning project activities.

Fourth, top-level decisionmakers in government and donor agencies frequently do not want to be bothered by the day-to-day concerns of program management. After initial program planning and approval, their interest shifts to monitoring project outputs.

Summative or product evaluation is conducted in most international technical assistance programs as a requirement of the funding or sponsoring group. Plans for the summative evaluation are generally developed and included in the initial project description at the time of the funding decision.

Product evaluation, in theory, has the largest audience of the various types of evaluation discussed here, but, in practice, often has the smallest. Project evaluation results potentially are of interest to the sponsor, the program implementors, and the various participant groups who were asked to invest their time, interest, or resources. These groups want to know what effects were achieved and what outcomes were attained. Often they seek confirmation of what they already believe about a program, based on their own experience with it.

A primary audience for product evaluation, however, should be the planners of future projects who may wish to replicate successful components of earlier programs. To the extent that replication occurs, evaluation contributes to a cumulative knowledge of international development strategy and can make a valuable contribution to subsequent project design efforts. The reason that product evaluation, in reality, has a small audience is because the replication of earlier projects is so infrequent.

Three reasons account for this tendency. First, social, economic, and political contexts differ so widely from country to country that, while general concepts may apply across settings, specific programs have limited application. Second, project design teams do much of their work in the field, away from easy access to the evaluation reports that might be of relevance and benefit to them. They typically work under time pressures that do not allow much time for them to search for relevant findings from previous product evaluations. Third, summative evaluation reports are sometimes written in a manner that undercuts their usefulness. Specifically, considerable attention is given to analyzing and reporting results, while little attention is given to describing processes of the program in enough detail to allow for an understanding of what really happened in the project.

III. THE NATURE OF THE EVALUATOR

Three issues to be considered in choosing an evaluator are whether the evaluator should be a member of the project staff, a content expert, and/or a country expert. The decision as to whether the evaluator should be internal or external to the project staff depends on the role the evaluation is to play within the project. Context and input evaluations are conducted before the project has been designed, so the evaluator is necessarily external to the project. The emphasis in process evaluation, however, is to provide ongoing feedback to the project manager about the functioning of the project. This requires a detailed knowledge of project activities, a sense of how day-to-day management decisions are made, and a close working relationship with project management. Often, this is best accomplished by having an evaluation expert as a member of the project staff.

Product evaluation answers a different type of question--whether a program should receive extended funding, be disseminated more widely, or adopted elsewhere. The use of an internal evaluator poses an obvious conflict of interest. As a member of a program staff, an internal evaluator may feel that reporting negative findings is disloyal to friends and colleagues on the staff. Further, the evaluator may be placed in the uncomfortable position of reporting results which, if accepted, could result in termination of the project and a loss of his/her own employment. Consequently, summative evaluation should be conducted by independent, external evaluators who do not stand to gain or lose personally from the evaluation results.

Within reasonable limits, the evaluator does not have to be an expert in the content of the program being evaluated. It is more important that the person have good evaluation skills, including the ability to seek content expertise from others when necessary. An evaluator untrained in social studies education would not presume to judge the appropriateness of the content coverage of a proposed curriculum. However, the evaluator should be able to identify that a content review is an appropriate activity, organize the team to review the materials, and present the results of that review.

At times, being a content specialist may even impede an evaluator's performance. This happens (a) when, as content experts, evaluators have such a strong bias about how the program or content should be presented that it intrudes on their objectivity or (b) when evaluators focus an evaluation to highlight their own area of expertise and fail to provide appropriate emphasis to other dimensions of the activity. For example, an expert in vocational education may be an inappropriate choice to do a context evaluation of the vocational/technical subsector, since many of the critical issues in the study will emerge from the articulation of vocational education with other sectors of the education system and from the analysis of labor supply and demand projections. The analytic skills required to do these tasks typically are not part of the training of vocational educators and the evaluator will be forced to make judgments concerning the relative priority for vocational/technical education vis-a-vis other EHR sectors.

This is not to say that background and experience in the content area under study are not useful, nor is it to suggest that all issues are best addressed by evaluators who have only a generalist's knowledge of the substantive content. If vocational educators are available who have both the professional objectivity and the broad analytic training required in a context analysis, their vocational background may be an advantage. Clearly, if school construction plans are being evaluated, engineering skills are a necessary, even if not a sufficient, requirement for a proper evaluation. The appropriate credentials of the evaluator will depend on the specific circumstances of the program to be evaluated. However, specific subject matter expertise frequently is secondary to the person's expertise in evaluation.

Another issue frequently of concern in selecting an evaluator is whether the evaluator should be a country expert. A knowledge of the issues in international development and a sensitivity to the social, cultural, economic, and political dimensions of the country are essential; specific country experience is not. However, adequate time for evaluators new to a particular country to learn about the society and culture should be incorporated in the time allowed for planning and conducting the evaluation.

There are few additional guiding principles about who should conduct evaluation. The person should be appropriately trained as an evaluator, be able to establish a productive working relationship with government and project personnel, and be sensitive to constraints imposed by the setting in which the project operates. Programs in developing countries place heavy demands in all three of these areas. First, international program evaluations are frequently conducted by people well versed in social science research methodology, but with little background or experience in evaluation. This occurs largely because program personnel do not understand evaluation or its distinction from research.

Second, large-scale educational programs in developing countries operate in a complex political environment, marked by sharp differences in the political and economic power wielded by the interest groups associated with a particular program. The evaluator must be politically sophisticated and interpersonally skilled to avoid the subtle manipulation and pressure that may be exerted by powerful groups and to recognize and adapt when such manipulation and pressure cannot be avoided.

Third, the social, cultural, and economic context in which a program operates places demands on a program that the evaluator needs to understand. The issues may be as straightforward as understanding the distribution problems posed by a fuel shortage or as subtle as noticing the difference in the number of times classroom teachers call on female versus male students.

CHAPTER FOUR

CRITERIA, STANDARDS, AND INDICATORS

I. INTRODUCTION

This section argues that basic to evaluation practice is the distinction among criteria, standards, and indicators. Criteria are the characteristics of a program regarded as important bases for evaluating that program. Standards answer the question "How much of the important criterion is enough?" Indicators refer to the measures used to collect data regarding performance on the valued criteria.* Two problems confront current thinking about evaluation: (1) People frequently fail to distinguish among criteria, standards, and indicators; and, (2) large-scale social and educational programs serve many audiences and have multiple stakeholders. The various stakeholders in a program often disagree on the importance of criteria, the appropriateness of particular standards, or the relevance of indicators. These considerations underlie the need to incorporate multiple criteria and employ multiple measures in the conduct of educational program evaluation. Further, the evaluation should serve to clarify and expose the standards that key stakeholders use in formulating their own judgments of program worth.

Criteria are the characteristics of a program which are regarded as relevant and important bases for evaluating that program. For example, student achievement might be used as the criterion for judging the success of new instructional materials. The teacher attrition rate might be the criterion on which a national teacher salary supplement program is evaluated. Community utilization might be a criterion for evaluating an agricultural extension program.

Criteria are an expression of what people value about a program. These valuations are grounded in beliefs, personal experience, the experience of others, and the results of theory driven research. Criteria cannot be logically deduced; their distinguishing

The distinction among criteria, standards, and indicators can be traced to the work of Dewey. A more current discussion of the distinction in the context of program evaluation is provided by Moritz Johnson. Portions of the present discussion are drawn from his article on "Evaluation Reflections: The Locus of Value Judgements in Educational Program Evaluation," which appeared in Studies in Educational Evaluation, 5, 1979, 109-122. The discussion of criteria draws from the work of R. Sadler, "The origins and functions of evaluative criteria," Educational Theory, 1985, 35, 3, 285-297. See also G. Glass, "Standards and Criteria," Journal of Educational Measurement, 15, 4, 1978, 237-261.

characteristic is that they must be chosen. They represent judgments about what is important about a program.

Individuals may disagree about the criteria they think are important. In part, this occurs because people differ in their personal experiences, in their familiarity with the proposed and with similar programs, and in their knowledge of relevant research. Further, they disagree because any single educational program may have many different outcomes, each of which holds different consequences for the key groups involved in or affected by the program. A common criticism of an evaluation is to claim that the study was conducted adequately but that the results are irrelevant because the wrong criteria were investigated. A key issue in designing an evaluation is specifying the criteria on which the program will be judged.

Sadler (1985) suggests that criteria can be organized into a hierarchical structure. Any given criterion can be expressed either as a component of some higher level criterion or as inclusive of some lower level criterion. That is, a criterion is given content by specifying it in terms of a number of simpler criteria. Specifying the simpler criteria (lower in the hierarchy) is a matter of interpretation and semantics. For example, if appropriateness is a criterion for choosing a textbook, it might be specified in component terms that include coverage of content matter, depth of treatment, reading level, and pertinence of examples. Connection to higher level criteria is developed by asking why the criterion (e.g., appropriateness) is thought to be worthwhile.

Specifying a criterion in lower level criteria is useful in clarifying and communicating what is important in a project. Specifying criteria also simplifies the process of judgment. By restricting the dimensions on which comparisons are possible, a value claim is easier to establish.

Standards refer to the desired level or quality of the criteria. For example, what level of student achievement in mathematics must be observed for new instructional materials in that subject to be regarded as successful? What reduction in teacher attrition will be accepted as evidence that the salary incentive program was effective? What rate of community utilization of the information provided through the agricultural extension project is necessary to justify continuing the program?

Standards, like criteria, are issues of judgment. The required amount of an important program characteristic may vary considerably by individual or by groups. Even within a group, standards held for a program may vary over time and location. Standards can be either relative (e.g., students with the new material performed better than students using the old materials) or absolute (e.g., students successfully mastered 85% of the content).

Indicators are the measuring instruments used to collect data about performance on the criteria specified as important. Examples of measuring instruments include:

- questionnaires
- observations
- interviews
- norm-referenced tests
- criterion-referenced tests
- delphi technique
- Q sort
- expert review
- checklists

Selection of measuring instruments is an important step in conducting evaluation studies, because the choice of instruments is the final point at which the evaluator determines how the criteria for the evaluation are to be defined operationally. Three factors bear on the choice of measures.

First, different types of data can be collected that address the same criteria. Indeed, several types of measuring instruments sometimes are needed to assess peoples' performance on a single criterion adequately. For example, the achievement of ministry employees in a management training course might be assessed either by a written test covering what the trainees learned, by rating their performance in a simulation exercise that requires trainees to demonstrate what they learned, or through observation of on-the-job performance.

Second, educational programs often have multiple criteria of success which, in turn, require that several types of measures be used. For example, the success of a textbook project may rest on both the achievement of students using the textbooks and the attitude of the teachers toward the new materials.

Third, the quality of the data gathered in a evaluation study is no better than the quality of the measuring instruments employed. Criteria for instrument selection include:

- Validity - Does the instrument measure what it is supposed to measure?
- Reliability - Does the instrument yield a consistent measurement?
- Usability - Can the measure be used meaningfully within the evaluation setting?
- Appropriateness - Is the measure appropriate for use with the respondents?
- Availability - Can the measuring instrument be obtained in sufficient quantity within the time frame available?

Procedures for the selection and development of measurement instruments are discussed widely in the evaluation literature.* The reader is referred to those sources for further consideration of general issues in instrument design and selection. However, the cross-cultural use of measurement instruments raises a special set of technical concerns for evaluators working in international settings. The appropriateness of the measuring instruments often is limited by the cultural setting.

II. CONFUSION AMONG CRITERIA, STANDARDS, AND INDICATORS

Program personnel often fail to distinguish between criteria and standards and may use the terms interchangeably out of ignorance or carelessness. Although one frequently hears that something "meets" a criterion, it is standards that are met, not criteria. Similarly, standards can be "high" or "low," as criteria are "appropriate" or "inappropriate." While criteria and standards are related, their determination requires two quite separate decisions.

Similarly, the distinction between criteria and indicators is sometimes overlooked. This distinction is particularly useful because it reminds us that there often are several alternative indicators of the same criterion. The distinction is overlooked when personnel become committed to particular measures and talk as if the performance of these measures is the most (or only) important outcome of their program. They may treat the measuring instrument as if it were the criterion. For example, a teacher or school administrator may talk about test performance as if that was what the schooling experience was designed to produce, whereas test performance is only one indicator of student achievement.

In international settings, the criteria/indicators distinction is particularly important, since the circumstances under which the data must be collected may require flexibility in the types of measures that can be used.

A particularly useful discussion of the problems in conceptualizing and measuring implementation is provided by Fullan, M. and A. Pomfret, "Research on Curriculum and Instruction Implementation," Review of Educational Research, 47, 1, 1977, 335-397. See also Hall, G. and S.A. Loucks, "A Development Model for Determining Whether the Treatment Really is Implemented," Research and Development Center for Teacher Education, University of Texas at Austin, 1976. A discussion of implementation as a criterion in evaluation is provided by Dobson, D. and T.J. Cook, "Avoiding Type III Errors in Program Evaluation: Results From a Field Experiment," Evaluation and Program Planning, 3, 1980, 269-276. In Dobson and Cook's definition, a Type III error is evaluating something that never occurred.

Disagreement among Stakeholders. Large scale educational programs serve many audiences and have multiple stakeholders. Stakeholders are people and groups interested in and affected by a program. Stakeholders are people who think they have something to gain or lose from a program.

It is common for the failure of an otherwise well-designed program to be attributed to the fact that a key group did not participate in a manner required for project success. The designers did not always fail to attempt to incorporate the keygroup into the design process. Rather, group members may not have understood the program design or their role in the program. They may have believed that the program's success eventually would work to their disadvantage, perhaps resulting in additional work, undesired reassignment, a reduction in income, or a loss of their employment. The non-participants may have seen the program as a threat to some other activity or program to which they were committed or, internally, group members may have differed on what they regarded as important about the program.

Even when all relevant groups support a program, they may do so for quite different reasons and value the program on quite different criteria. For example, teachers may support an inservice teacher training program for the opportunity it gives them for promotion or additional pay. Support from parents comes because they believe it will improve instruction and enhance their child's education and eventual employment opportunities. The Ministry of Education may support a program because of a belief that it will increase teacher retention. In fact, the program may do all these things. However, evaluating the program only in terms of the teacher retention rate would fail to explain why the program elicited the pluralistic support that contributed to the program's successful implementation.

One of the most important issues in evaluation design is the manner in which an evaluation takes into account the views and judgments of the multiple individuals and groups involved with a project. If evaluation is conceived only as a determination of whether intended outcomes occurred, the experiences of certain groups involved in or affected by the program may be overlooked and their judgments about the worth of the program ignored.

Five reasons help account for why education programs, in particular, involve so many stakeholder groups. First, the social demand for education by parents is intense. Parents have high (and often unrealistic) expectations of the benefits that will accrue to their children from educational participation. In most developing countries, education is seen as the primary route to upward social and economic mobility. Consequently, the education system has been characterized by explosive growth in school enrollments. This pattern of growth may be partially economic in origin, a result of emerging occupational structures (Foster, 1985). However, the pattern is exacerbated by the demographic phenomenon of an increasing percent of the population being of school age and an increasing percent of school age children wanting to attend formal schooling.

Second, education tends to command a disproportionate share of national resources. Governments have felt compelled to respond to the intense social demand for education, partly from a belief in education's ability to stimulate aggregate economic development, partly to avoid frustrating the aspirations of large segments of the population. This has led to large investments in education and a potentially disproportionate emphasis on the education sector at the expense of other development activities.

Third, the outputs of the education system can constrain or facilitate the success of other sectors of the economy. Government planning is based on the projected availability of personnel with the skills necessary to implement the plans. Inadequate training, in either quality or a quantity, undercuts the effectiveness of other sectors that depend on the input of trained personnel.

Fourth, in most developing countries, education operates as a centralized activity that has high visibility and presence at the local level. Particularly in rural areas, schools are one of the more tangible aspects of central government activity in the locality. The activities of the school are observed and discussed widely by parents and community members. Indeed, schools often play a larger social and cultural function in the community and are a major factor in modern nation- building activities.

Fifth, while most children in developing countries are still first generation school attenders, most officials of government, business, and industry have completed at least some intermediate or higher level of schooling. These officials have experienced the status, income, and employment benefits to be derived from attainment within the formal education system and, on the basis of their personal experience, tend to have strong beliefs about the appropriate structure and content of education.

These social and political dynamics help explain why education programs operate within such a complex network of interest groups. The most obvious impact of these dynamics on evaluation is to politicize the process. Stakeholder groups differ in their purposes for supporting a program and in the types of decisions they must make about the program. They differ in the criteria they use, the standards they apply, and the types of data they believe are relevant and credible as evidence of program success.

In general, the number of stakeholders increases as a program becomes more complex and as resources associated with a program increase. The increase in resources utilizes the attention of three groups: those people who stand to gain by controlling and/or sharing in the resources, those advocating alternative uses for the funds, and those who are required to supply the resources.

One of the key steps in designing an evaluation is to identify the key stakeholders. In all programs, stakeholders would include the groups participating directly in the program. For an education program, these might include Ministry of Education personnel, regional or district administrators, school principals, teachers, students,

parents, project personnel, and funding source personnel. Each group experiences the program in a different way. Each may seek different ends from their participation.

Evaluation must also attend to other, less proximate, groups of stakeholders--those who are less involved in the project activities, but who are affected nonetheless. These include: (a) other agencies that might have been recipients of the funds committed to the program or that would expect to receive funds if the program being evaluated is discontinued, (b) groups that will hire graduates of the program, and (c) groups that may want to adopt portions of the program if it is deemed successful. It is essential to consider the views of the various stakeholders; the decisions they make about their own participation and support may influence program success as much as the decisions of the program administrators.

Disagreements among stakeholders over the appropriateness of the criteria on which to judge program success and the necessary levels of attainment stem from five sources:

- their proximity to and extent of participation in the program;
- their knowledge about the program--how it is designed, how it operates, the multiple impacts it may have;
- the outcomes in which they are interested;
- the extent to which they believe the program has consequences for them; and
- their power to influence program decisions.

These issues, then, can be used to help identify the key stakeholder groups in a program evaluation.

It may not be practical to consider the views of all stakeholder groups in conducting an evaluation. Which groups to include is a decision the evaluator makes, usually in discussion with the client. However, since clients are often associated with project management and have their own special interests to protect, the evaluator usually carries the major responsibility for designing the evaluative study in a manner that is responsive to a wider set of stakeholders' interests.

III. CRITERIA FOR EVALUATION OF INTERNATIONAL TECHNICAL ASSISTANCE PROGRAMS

Four key points summarize the preceeding discussion about criteria:

1. Criteria are the dimensions of worth on which a program or activity is evaluated.

2. A major source of confusion in evaluation occurs when people fail to distinguish among criteria, standards, and indicators.

3. The various stakeholders involved with a program may differ in the criteria they believe are the relevant bases for evaluation.

4. Stakeholders may differ in the clarity and explicitness with which they hold the same criterion.

Often evaluation has been conducted to determine the extent to which intended outcomes were achieved. While this is an important and widely shared criterion, sole reliance on goal attainment for the evaluation criterion seriously limits the utility and relevance of the evaluation activity. Further, evaluation can play an important role even before any explicit program goals are achieved.

This section discusses six criteria widely applicable to evaluations of programs to enhance educational efficiency in international settings. They are organized in three categories according to the aspect of program planning and implementation to which they apply. Criteria concerned with intents are relevant during context and input evaluations. Intents refer to peoples' plans for the structure, operation, and outcomes of a program. The criterion for evaluating intents is logical consistency, i.e., is it reasonable, given available information about the environment, that the proposed inputs, organized and operated in the stated manner, will lead to the anticipated outcomes?

Once a program has moved from the planning to the implementation stage, the appropriate evaluation criteria are those concerned with process. Two types will be discussed: the level of program implementation and process criteria.

As the program yields intermediate and final outcomes, a variety of specific project effects and impacts serve as criteria, depending on their value to some stakeholder group. Two which are relevant to all projects have already been considered in this paper: effectiveness and efficiency. In addition, two categories of project impacts, intended outcomes and unanticipated outcomes, will be discussed.

Logical Analysis. Evaluations that occur before a program is implemented rely on logical analysis to assess the extent of the logical contingencies among proposed inputs, intended processes, and desired outputs. The task of logical analysis is to identify the links between intended outcomes and the particular antecedent conditions and instructional transactions on which they are contingent. The data for a logical analysis consists of a full understanding of program intents, program resources, the context in which the program will be operating, and the management, administrative, and decision structure of the program. To test the logic of an educational contingency,

the evaluator relies on previous experience, the experience of others, and research in similar settings. In conducting a logical analysis, the evaluator compares probable and possible system performance and analyzes possible causes of discrepancies between actualities and intention.

Evaluation using logically derived criteria is particularly important in international settings. First, international technical assistance programs tend to be expensive. Inherently, the costs tend to be front-loaded in the project, with many costs incurred before the program is sufficiently operational in the field for unanticipated problems in the logic of the planning to surface. Early identification of logical inconsistencies can save substantial resources.

Secondly, pressures from the multiple interest groups that influence project design may succeed in including components that serve the interest of a particular group but not necessarily the program. Sometimes these pressures are applied subtly enough so that concessions and modifications are incremental--program inputs, processes, and outputs may no longer be aligned. Project planners may not realize the degree to which a program design has been altered to accommodate interest groups.

However, logical analysis also has particular pitfalls in international settings. To assess the logical contingencies between context, anticipated inputs, proposed processes, and intended outcomes requires a substantial knowledge of both the program components and the context in which the program will be operating. Evaluators unfamiliar with the culture, the infrastructure, or the educational system of the country for which the program is being designed tend to make serious miscalculations about appropriate time schedules, the cultural reactivity of the program, the attractiveness of incentives used within the program, and the infrastructure constraints on communication and distribution activities.

The other risk is that individuals who know the culture and local situation well enough to provide a logical analysis may have strong, and often implicit, biases about the program and/or the setting which influence their judgment. The logical analysis is finally the responsibility of the evaluator who may represent an additional interest group. These points underscore the importance of the person selected to do the evaluation. The evaluator must have sufficient knowledge of the country and the culture while being perceived as independent and fair in approaching the study. As was discussed earlier, the criteria for selection of an evaluator includes project familiarity, context familiarity, objectivity, and technical expertise.

Level of Implementation. Evaluations of large-scale educational programs frequently find no immediately observable effects. The reasons for this may be that the educational treatment to improve efficiency never occurred, was inappropriately delivered, or the sampling process was highly idiosyncratic. As a result, potentially effective programs may be dismantled due to negative results derived not from the

ineffectiveness of the program concept, but from a failure to consider whether the treatment was ever delivered at all.

Level of implementation refers to the extent to which a program is actually implemented or used. It differs from the decision to adopt the program and is not synonomous with planned or intended use. Implementation is more that an extension of planning; it is a phenomenon in its own right.

In program evaluation, however, level of implementation as an evaluation criterion is often overlooked. Once a program is planned and approved, the monitoring of program operation typically is relegated to program staff. The interest of higher-level policymakers tends to shift toward the monitoring of outcomes. Often, there is little curiosity about what happens to the program between the time it is planned and approved and the time the consequences become evident. Yet, program outcomes depend on how the program actually operates in practice. This, in turn, depends on the daily activities of those organizational members in charge of implementing it--as managers, participants, and support personnel.

Level of implementation is a criterion of particular relevance in the evaluation of educational programs in developing countries for two reasons:

1. Many of the most important outputs of large-scale programs are long-term. Yet, program sponsors frequently require that evaluations be conducted before it is reasonable to expect that the most meaningful outputs would be observable. To show that the program is being implemented as planned offers a more meaningful alternative in the short-term than looking for outputs that would not yet be expected to have matured.

2. Fielding new programs in developing country settings poses particularly severe problems of implementation--unreliable communications, lack of transportation, difficulties in securing materials and supplies, and often, limited staff to monitor program activities. A real risk in evaluations in developing countries is that considerable time and money is expended evaluating outputs of project activities when, in fact, the activities have not yet occurred.

Evaluators give considerable attention to strong evaluation designs to rule out rival interpretations of data but often fail to observe key issues in implementation. Where efforts to evaluate implementation do occur, the efforts are often directed to monitoring program inputs rather than processes. Monitoring inputs addresses the extent to which funds, personnel, and material are being committed on schedule. The importance attached to this type of monitoring activity derives from three factors.

First, the outcomes desired of many education programs are long- term (including employment, occupational choice, lifetime income, and effects on national goals of

development and equity). In the interim, there is a widespread willingness to accept short-term outputs as proxies for outcomes and inputs as proxies for expected outputs. This willingness can be attributed to sponsors placing great faith in the logical consistency of their own planning, in which they linked inputs to outputs and outputs to outcomes through their project design. In addition, inputs are relatively easy to quantify as long as appropriate records are maintained. Further, it is not uncommon for sponsors to evaluate program managers on the extent to which financial commitment levels stay on schedule. Hence, program administrators may exercise pressure for a program to stay on schedule, even when the addition of new inputs may be premature or out of phase with other project activities.

While the monitoring of inputs is important, the input of resources to a program is only a partial indicator of implementation. Further, it is frequently a poor indicator of more substantive issues of how well a program actually is functioning.

Investigating implementation directly addresses some of the more important issues of whether a new program succeeds. Fullan and Pomfret (1977) point out that many of the problems associated with introducing curricular innovations do not reside in the actual development or production of the curricular materials, in getting people to try the innovation, or even in getting them to use the materials in a certain way. Rather, the main problem appears to be that the curriculum change usually necessitates certain organizational changes in the roles and relationships of organizational members and changes in the way organizational members behave toward each other. However, the organizational implications of new educational programs are rarely explicit in the plans.

An evaluation should address at least three issues when examining program implementation.

1. To what extent was the program implemented? The concern here is with the fidelity between what was planned and the program as it presently exists.

2. What were the impediments to implementation?

3. What adaptations and changes occurred in the implementation process that should inform further program planning or dissemination?

Models for conceptualizing the fidelity of implementation and the factors influencing degree of implementation are described in social science literature and will

not be discussed in detail here.* However, evaluators should attend to four particularly important determinants of implementation:

1. Explicitness--the extent to which people understand the program.

2. Complexity--the degree of difficulty users have in applying the program.

3. Degree of change--how different the program is from existing programs and practices.

4. Incentives--what encouragements are available to users for correct implementation.

The main data collection methods used in evaluations of implementation have been observation techniques, focused interviews, questionnaires, and content analysis of key documents and program materials. For these, observation techniques offer the most rigorous measurement of the behavioral fidelity of implementation to the original plans. However, observations are subject to several problems. One is the impact of the observer on the performance of the program user. This is particularly acute in international settings when programs are implemented in rural areas in which the appearance of an outsider may itself be disruptive. Also, observation is labor intensive, expensive, and sometimes unfeasible given transportation difficulties, fuel shortages, and language problems. Finally, observation may tap only the mechanical use of the program and not assess other dimensions, such as the users' understanding of the philosophy or strategy of the program. Focused interviews can help solve this third problem and are especially effective in uncovering why users made modifications to the proposed practices.

Process Criteria. One set of evaluative criteria concerns the process and intermediate outcomes associated with the actual delivery of a program. These outcomes are not the

A particularly useful discussion of the problems in conceptualizing and measuring implementation is provided by Fullan, M. and A. Pomfret, "Research on Curriculum and Instruction Implementation," Review of Educational Research, 47, 1, 1977, 335-397. See also Hall, G. and S.A. Loucks, "A Development Model for Determining Whether the Treatment Really is Implemented," Research and Development Center for Teacher Education, University of Texas at Austin, 1976. A discussion of implementation as a criterion in evaluation is provided by Dobson, D. and T.J. Cook, "Avoiding Type III Errors in Program Evaluation: Results From a Field Experiment," Evaluation and Program Planning, 3, 1980, 269-276. In Dobson and Cook's definition, a Type III error is evaluating something that never occurred.

primary outputs of the project, but are a consequence of the transactions that occur as a program is implemented. Examples include the rapport and goodwill between project personnel and trainees, host government officials, and the sponsoring agency staff; the attitudes and feelings of participants toward the program; and tangible accomplishments that may be incremental steps toward some other goal but which have importance in and of themselves.

Process criteria play an important role in international technical assistance programs for three reasons. First, education programs have a poorly defined process component. The planning activity typically concentrates on the identification of inputs, design of the overall program structure, and specification of intended outputs and outcomes, but it seldom addresses the specific types of transactions and intermediate processes that must characterize program implementation. Indeed, process variables are given the least attention in the planning process, often, under the assumption that these are the responsibility and purview of the team implementing the program in the field.

Second, many significant program impacts emerge only as long-term outcomes--too late to contribute to some program decisions. For example, refunding decisions on a five year project may begin in project years three and four--long before the outcomes or even outputs of the first funding cycle might be evident or could be assessed. Process phenomena serve as proxies of longer-term outcomes that have not yet been achieved. Further, they shape the probabilities that the desired outcomes will be achieved.

Third, in the complex interest group structure surrounding educational programs, the intermediate process and output components of a program are often the most important components to the various interest groups. Indeed, some stakeholders of a project may have little interest in the stated purposes of a program or its longer-term outcomes, but they support the project to secure intermediate or tangential benefits they believe will follow from program implementation. The intermediate project effects may be the primary outcomes some stakeholders hope to achieve. For example, specific stakeholder group interests might include:

- Improved bilateral relations as an outcome of good project/government relationships.

- Conditions precedent (prior conditions a recipient government must meet to receive funding), in which the leverage offered by the promise of funding facilitates other decisions by a host government that might be only tangentially related to the direct purpose of a project.

- Cash flow and foreign exchange benefits provided by an education project to a country suffering serious economic problems.

- English language training received as a component of a more specific skill training program. The objective of the language training is to prepare

participants for entry to a training program taught by expatriates, but the English language training itself is seen by participants as a means to enhance their opportunities.

The difficulty for evaluators is that process phenomena and measures often are not explicitly stated in the original program design. This may be because those effects are tangential or unrelated to the direct purposes of the project; are held by interest groups not directly involved in delivery of the program itself; were never thoughtfully considered and remain unarticulated; or were unanticipated and, therefore, not deemed important until they occurred and their consequences became evident. As a result, there is likely to be little consensus about the importance of these intermediate effects, since they are not explicit and may never have been discussed.

Generally, the evaluator must take responsibility for identifying and articulating the process criteria they are appropriate as bases for program evaluation. This involves eliciting from key groups, usually through interviews, what they regard as the important processes and effects of the project. However, their effort may meet with resistance. Identifying and reaching the appropriate stakeholders often poses a problem, exacerbated in international settings by language and logistical barriers. Even when contacted, stakeholders may not be able to articulate the processes and intermediate outcomes they value. This is not to say that they would not recognize or value those effects when they occur. Also, program administrators may resist the evaluator's efforts in this direction because the activity suggests that the evaluation may be based on criteria other than those the program was designed to achieve, perhaps even beyond the influence the program administrators. Work by Stake (1967) offers a useful framework for conceptualizing types of process criteria. He suggests that evaluators should collect the beliefs and judgments of key groups regarding both the transactions and intermediate outcomes they believed the program would achieve prior to the program beginning and which transactions and intermediate effects those groups believe are actually occurring.

Program Outputs. Outputs refer to the impact and effects occurring as a result of the program. Specific output criteria on which a program is judged might include student acquisition of specific knowledge, skills, and abilities; development of desired attitudes, behaviors and workstyles; and higher productivity. At an institutional level, it might include the development of new organizational skills, improved management systems, or enhanced fiscal capacity.

Effectiveness and Efficiency. The specific outputs will vary with the nature and purpose of the program. However, two supraordinate output criteria that apply across all projects are effectiveness and efficiency. Effectiveness refers to the amount of a desired output achieved relative to the quantity of input. As was noted earlier, effectiveness and efficiency are measures of a program's goal attainment and not goals in and of themselves. They can be goals in a organizational sense only if one has a project that is an attempt to improve effectiveness and efficiency in the use of existing

resources. In this case, the efficiency program or project does not add resources to the existing educational system, and it accepts system goals as given.

Intended and Unintended Outputs. Intended outputs are those outputs which have served as objectives of the project. Their attainment is the reason the project was funded. They are the targets that a program seeks to achieve, the effects that someone thinks are most worthy. The extent to which the intended outputs are achieved is a necessary criterion for a program evaluation.

It is not, however, a sufficient criterion. All programs have outcomes, impacts, and effects beyond those that planners anticipate. Unintended outputs (or externalities) are those effects of the program which were not part of the direct rationale and purpose that supported the initial project funding. Unintended outcomes are not necessarily undesirable--their impact may be either positive or negative.

On one hand, intended outputs are the easiest evaluative criteria to identify because they emerge directly from the planning process. The purpose and objectives of a program define the criteria on which the program will be held accountable. Clarity about the important criteria allows the collection of appropriate baseline data, selection of appropriate measures, and the development of an appropriate evaluation design.

In practice, the determination of whether a program met its objectives is seldom that easy. The evaluator encounters the following problems. First, intended outputs often are stated in nebulous terms. Second, intended outputs may not follow reasonably from the project activities due to a failure in the logic of the project design. Third, the purposes of a program often shift as the program operates.

Ambiguous statements of intended outputs often occur when program objectives are stated at high levels of generality. Sometimes this is intentional--program objectives are stated in general terms as a means of securing wider consensus among potential stakeholders. At other times, the generality merely reflects a lack of clarity about what the program intends. This is not uncommon in complex programs which seek to accomplish multiple outputs.

The more serious problem is when the intended outputs do not align with the program offered. This misspecification of outputs often reflects a confusion over educational processes that can be traced back to the initial program design process. Program planners may overstate the potential impacts of their programs. For example, planners may claim that inservice training of central ministry staff will lead to increased achievement of students in schools, even though no direct or immediate link exists between improved central ministry staff performance and student learning. These exaggerated claims are due sometimes to the over-zealousness of planners hoping to secure support for their projects. At other times, they are the result of poor planning in which the links among project components have not been elaborated clearly.

This problem is illustrated in the scheme presented in Figure One. A common implementation strategy in large-scale development projects is to concentrate initially on developing the infrastructure within the sector of interest (e.g., education) and only then move project activities to lower levels of the organizational structure. This means, for example, that project activities might be concentrated initially on training and collaborative interaction with central ministry personnel. As trained personnel are available at the ministry level, project activities shift to working with regional personnel and, eventually, to working with school personnel directly. The rationale for this tiered approach is twofold: (a) it involves local staff in subsequent training which, in turn, (b) helps to institutionalize the program. These levels form the rows in Figure One.

In their simplest form, project designs usually specify the needed inputs, anticipated processes, and intended outputs of a project. These typically are specified for each of the levels at which the project proposes to work. These components (inputs, processes, outputs) represent the categories across the top of Figure One.

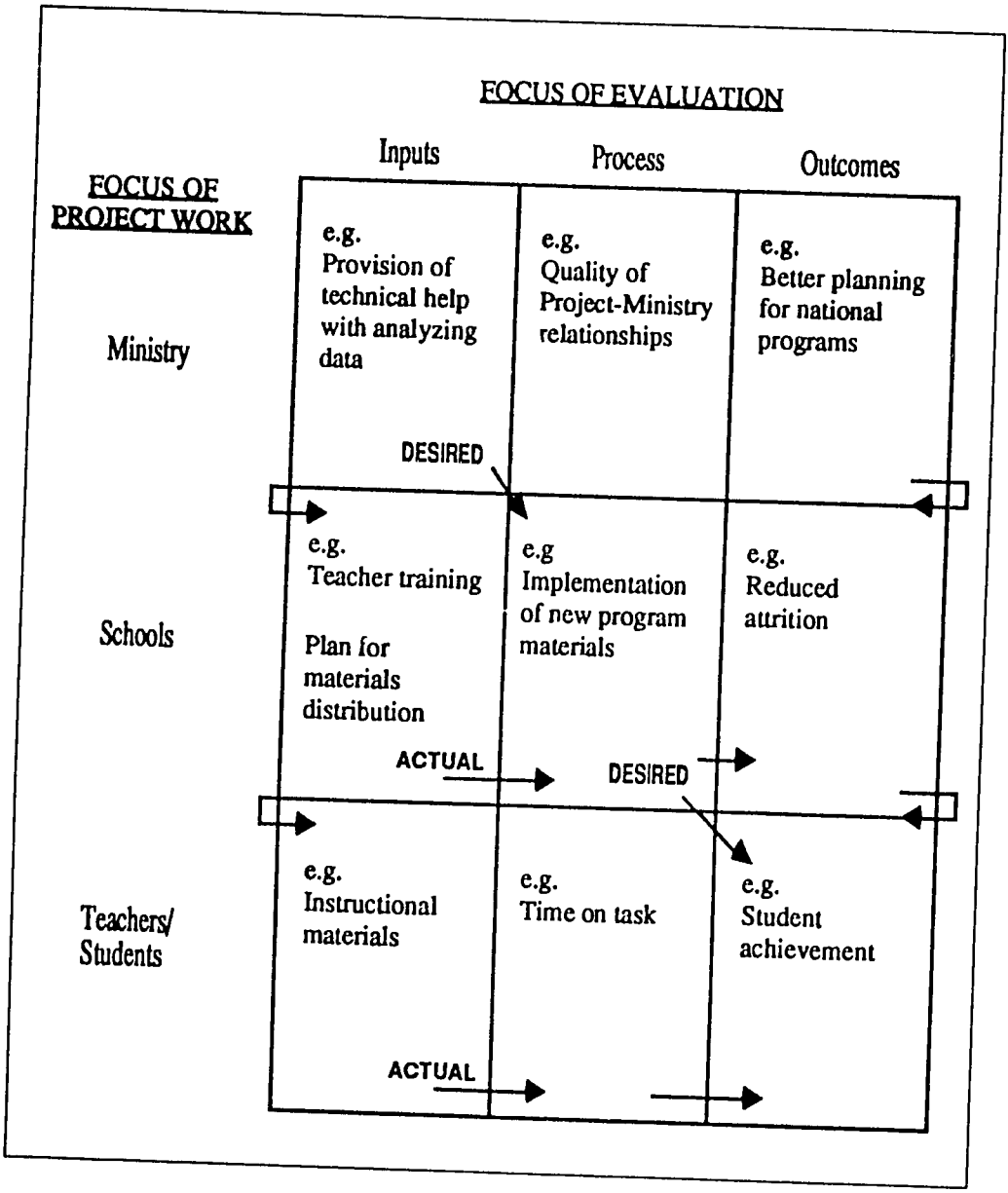
The desire of many planners is to try to make claims that move diagonally across Figure One. Inputs at the ministry level are expected to impact on processes at the school level which, in turn, are expected to be evidenced in increased levels of student achievement. Such logic may be specious, however, since the reform of educational programs rarely operates in this manner. Rather, inputs at one level of aggregation operate on the next level only to the extent that they result in outputs at that same level which then serve as inputs into the next level. International development activities would be expected to operate vertically and horizontally across Figure One, but rarely diagonally.

Shifts in program purposes and intended outputs that occur during implementation frequently are necessary and appropriate modifications made in response to the changing setting in which the program operates. Nonetheless, these shifts, though widely understood by the field implementation teams, may never be recorded formally or communicated to the sponsor. The evaluator in search of evidence that addresses initially anticipated outputs may not understand that the shift to different objectives was intentional and may not be sure that the new objectives really represent the wishes of other key audiences.

Programs have many more effects than those specified by planners. Further, the comples and shifting nature of education and social programs leads to outputs that original planners may never have anticipated. Indeed, for some stakeholders, the worth of some programs rests more on the unintended outputs that occurred than on the effects the program initially was designed to accomplish. Unintended effects can be either positive or negative. However, as Weiss (1975) points out, in practice they are more often negative, because program planners trying to justify their program are more likely to have cited all of the possible positive results.

FIGURE ONE

Simplified Model of Desired Program Impacts



The problem for the evaluator is to identify which unintended outputs to monito and then do so in time to collect appropriate data. Three strategies are available. First as previously discussed, the evaluator should determine the outputs important to ke audiences of a project. Many of these groups will not have participated in projec planning or in choosing the "intended" outputs. They do, however, have beliefs abou the program outputs they value and seek to achieve through their association with th program.

Second, the evaluator needs to document shifts in program purpose and operatio that may affect the range of effects to be assessed. A logical analysis of these shift may suggest additional outputs that should be monitored.

Third, the use of multiple measures increases the range of possible outputs that ca be assessed. Evaluators should select measures that assess a range of possible effec beyond those formally stated in planning documents. The use of multiple measure should represent a conscious effort of the evaluator to detect program outputs beyon those specifically intended.

CHAPTER FIVE

THE EVALUATION PROCESS

I. STEPS IN CONDUCTING AN EVALUATION

Figure Two presents the eleven steps in conducting an evaluation. These steps represent key decisions the evaluator needs to make in the process of an evaluation design. Steps 1-6 address the social context in which the evaluation is to be conducted. Steps 7-11, while also concerned with the social context, turn to more technical issues of design, such as data collection, analysis, and interpretation. The evaluation process in Figure Two does not specify a particular point of view, evaluation design, or statistical procedure that should be employed. Rather, it identifies issues, both political and technical, that will need to be addressed within a larger set of decisions that face the evaluator.

The steps are not necessarily sequential--several of these decisions may be addressed simultaneously or in a different order. However, they are presented in the general order in which they normally would be encountered. While the issues raised at each step may be satisfied by any one of several answers, the position the evaluator takes at each point has consequences and implications for the choices the evaluator makes at other points.

A. Formulation of a Point of View

As discussed earlier, there are numerous evaluation models, each based on a slightly different view of what it means to evaluate. These differing views of evaluation are grounded in philosophies and beliefs about evaluation. However, no one point of view is necessarily the best or the most appropriate across all circumstances. Indeed, an evaluator may employ different evaluation models at different times or may operate from a "hybrid" model--a combination of views gleaned from several of the formal models.

The point of view that the evaluator adopts about the meaning of evaluation will shape answers made in response to many of the subsequent decisions encountered in the design and conduct of the study. For example, is evaluation a formal judgment of program worth or is it only a determination of whether intended program outcomes occurred? Is the worth of a program a judgment made by the evaluator, the program decisionmakers, or the audiences served by the program? Answers to these questions shape decisions about what types of data to collect and from whom it should be collected.

FIGURE TWO

Eleven Steps in Conducting an Evaluation

1. Formulate a point of view	What does it mean to "evaluate" the Program?
2. Identify purpose (rationale)	Why is this evaluation being done? Is this particular evaluation part of a larger issue being considered?
3. Identify client	Who is asking me to do it? To whom do I report the results?
4. Identify audiences and sponsor	Who needs the information from this evaluation? The audiences for the evaluation may be different from the audience served by the program itself. The various audiences will have different information needs. Who is commissioning and paying for the study?
5. Identify resources and constraints	What materials, personnel, time and constraints and previously collected data, etc., are available to the evaluator? What constraints are identified?
6. Specify the evaluation question	What questions should this evaluation address? Questions addressing issues of importance to various audiences are essential in evaluation. However, not all issues identified as important can be addressed. The evaluator will have to decide the issues on which the evaluation will focus.
7. Formulate an evaluation design	What types of evidence should be collected? From whom should it be collected?
8. Select a data collection procedure	How should the evidence be collected?

9. Collect data	How can I ensure that the data is collected without bias? How can I ensure a sufficient response rate?
10. Analyze data	How will the data be analyzed? What is the simplest, clearest, and most appropriate procedure for analyzing the data?
11. Interpret and report results; follow-through with results	How will the data be reported? How can I report the results so the reader can understand them most easily? What can I do to help the clients understand the implications of the results for their situation?

B. Identification of the Purpose (Rationale) for the Evaluation

Evaluations are undertaken for many reasons, not all of them obvious, noble, or even appropriate. Some evaluation studies are undertaken for their public relations value; some are conducted only to comply with sponsors' requirements; some are conducted as an effort of one person or group to embarrass another; and some are conducted to help program planners improve their program or to help decisionmakers chose among program alternatives.

Early in the evaluation, it is important to identify why the study is being undertaken, whether the study stands alone or is part of a larger issue being considered, and what the client's motives are for undertaking the study. Understanding the client's motives for undertaking an evaluation may help the evaluator to identify the relevant stakeholders in an evaluation, the relationships among stakeholder groups, unarticulated issues which may influence conduct of the study, and the openness of different groups to the use of data and evidence in examining the program.

C. Identification of the Client, Sponsor, and Key Audiences

Who commissions, funds, and monitors an evaluation activity? For whom does the evaluator work? To whom is the final evaluation report submitted? Which other persons and groups have a legitimate right to see the evaluation results, and who decides? These are important questions to clarify early in a study. However, the answers to these questions are not always obvious or clear.

In international technical assistance projects, these issues may be particularly confusing for two reasons. First, many development programs are funded by one group, such as an international donor agency, but operated by another, such as the

relevant department in the local Ministry of Education. Similarly, the sponsor of an evaluation may be different from the office designated to organize and monitor the evaluation. Questions can arise as to where the evaluator's loyalties and responsibilities lie.

Second, the client and the various stakeholders of an evaluation often differ considerably in the power and authority they wield. The appropriate client of an evaluation may have to yield to the wishes of another person or group which exercises power or influence over the client.

Early in an evaluation, it is important for the evaluator to clearly identify the sponsor, the client, and the key audiences of the evaluation. The sponsor is the group that is paying for the study. The client refers to the person who actually commissions the study, the person by whom the evaluator is employed (within the context of the evaluation). In some evaluations the sponsor and the client are the same, for example, when the person or group that commissions a study also oversees it. However, as was noted above, in international development projects the sponsor is often a donor agency, while the client is the local administrative unit that receives the donor funds and actually conducts the project. The client, then, is represented by an administrative person, within the local unit, who works with the evaluator to specify the scope of work and to identify the key questions to be addressed by the study. Unless clearly stated otherwise, the final evaluation report is submitted to the client, who then typically would oversee its distribution to other interested parties.

In the complex social and political context of international development projects, it is important to be clear about which groups operate in the sponsor and the client roles. Clarity early in the evaluation can help protect both the evaluator and the client from competition among interested groups for control of the data, evaluation results, and public relations benefits resulting from an evaluation study. Clarity does not mean that these issues will not arise, but it does offer the evaluator some guidance in how to operate within the political sphere should the rights of competing groups become an issue.

Key audiences, or "stakeholders," are groups that have a special interest in the evaluation and groups that believe they might be affected by the conduct and/or the results of the evaluation. For example, stakeholders of an educational project may include program planners, participants, instructors, administrators, the funding agency, the relevant ministry, other ministries that would like to see the project or program resources allocated in different ways, perhaps community groups who believe that the project will affect them, and perhaps other donor agencies interested in the potential for funding similar projects. The stakeholders in an evaluation typically go far beyond the sponsor and client groups.

There are three reasons why an evaluation needs to identify the key stakeholders for an evaluation. First, the relevance and utility of an evaluation depends on the degree to

which it addresses the important questions of the groups served by a program. The information needs of these groups will vary. What the program administrators may need to know in order to improve a program (e.g., appropriateness of content; pacing of instruction) may be quite different from what a parent needs to know in deciding to allow additional children to participate in the program (e.g., individual achievement, hidden costs of participation). Knowing what decisions key groups involved in the project face can help the evaluator in formulating the evaluation question, specifying the types of data to be collected, and determining how the results of the study are to be reported.

Second, stakeholders may differ in the criteria they believe important, the standards they apply, and the types of measures they trust. If an evaluation is to be responsive to the information needs of the stakeholder groups, these differences need to be understood.

Third, stakeholder groups differ in the influence and power they wield. The evaluator needs to understand the social and political dynamics that operate among and within the groups that serve as key audiences. This understanding will ensure that the needs of less politically influential (but for other reasons important) groups are not overlooked, help forestall the possibility that audiences with particularly strong interests in the outcomes of the evaluation will try to influence the design or progress of the study inappropriately, and guide the evaluation team in operating across groups without inadvertently committing social or political errors.

Not all stakeholders of an evaluation are equally important or need to be considered to the same extent. The decision as to which stakeholder groups will receive attention is one best made by the evaluator after becoming familiar with the project. Factors influencing the evaluator's decision should include: (a) closeness of the group to the program, (b) severity of the consequences of program outcomes for the group, and (c) social justice.

D. Identification of Resources and Constraints

Clients frequently have little idea of the steps involved in evaluation activities, how long they take, or how much they cost. It often falls to the evaluator to develop cost estimates and timelines. In international settings, this can be particularly difficult since there often is greater uncertainty about the availability of data, the time and resources needed for data collection, the local capacity to analyze data, shipping and mailing time, and the extent of cooperation to be expected from both local and donor agencies.

Beyond the uncertainties over logistical support, evaluations in third world settings may require procedures and steps typically not included in evaluations in more developed countries. For example, sufficient time needs to be scheduled for translating and verifying the translation of policy and program documents relevant to the study, measurement instruments for use in data collection, and the intermediate and final

written evaluation reports to the client. Of particular concern is the translation of technical terminology in which words have a special technical meaning not apparent in the words themselves. Mistranslations can be embarrassing and undercut the credibility of the study when discovered.

Using a translator brings with it other risks. Unless the evaluator speaks the language of the country fluently, the translation process may operate as a filter, screening out nuances and more subtle insights that the client and key informants are trying to communicate.

When top ministry officials speak English, the evaluator may be led to believe translation is unnecessary and rely on direct conversations with those ministry officials. In this case, English fluency may inadvertently be used as a proxy for the importance of the ministry official. The evaluator may tend to rely on certain officials, not because they are best positioned to have an important perspective on the program, but because they are the ones who can communicate most clearly to the evaluator.

E. Specification of the Evaluation Question

A common criticism of evaluations is that they do not answer the important questions that people have about a project. When this happens, the evaluation enterprise is seen as irrelevant and as "missing the mark." Evaluations fail to address key issues for four reasons:

1. No evaluation question was formulated. The evaluator and/or client become infatuated with a particular instrument or procedure and fail to state clearly the evaluation question underlying their efforts.

2. The evaluator fails to clarify the key issues and decisions facing the client or formulates an inappropriate evaluation question.

3. The evaluation question seemed relevant at the beginning of the study, but the interests and concerns of the client shift over time.

4. The results of the evaluation raise questions that go beyond the scope of the original study, leaving the client feeling that the original evaluation question was inadequately formulated.

These problems often result from the failure of the evaluator to clarify and make explicit the important evaluation questions that face the client, to develop consensus and commitment of key audiences as to the evaluation questions to be investigated, or to clarify the limits of what an evaluation can and cannot tell the client about a program.

F. Formulation of an Evaluation Design

The evaluation design refers to the specific strategy associated with the selection of respondents and the procedures for collecting data. It is one of the most important considerations in conducting the study, since the design used largely determines the nature and strengths of the claims that can be drawn from the results of the study.

The purposes of the study and the uses to which the evaluation will be put should determine the study design. Strong causal claims are not a necessary part of evaluation, nor are results that generalize to groups beyond those involved in the project at hand. What is important are designs that yield relevant and sufficient data on the worth of the project being evaluated, that allow the relevant stakeholders of a program to make more informed decisions about the program under study.

In designing a study, the differences between evaluation and research must be kept clearly in mind. Two distinctions are of particular importance.

1. Strong causal claims are not a necessary part of evaluation; claims about the worth and utility of a program are. The purpose of evaluation is to support judgments about the worth of a program based on criteria deemed important. Rigorous claims about the extent to which the treatment caused the observed effects is frequently a desired outcome of an evaluation study. However, evaluations address important and meaningful issues beyond the causal relationships between treatment and effect. Indeed, it is quite possible that a program in which the treatment results in the intended effect may still be judged of little or no value by key stakeholders. This happens, for example, when a program yields unintended negative outputs in addition to the desired outputs, or, when key stakeholders value several outputs and differ in their levels of commitment to those specified as most important in the evaluation.

2. Generalization of evaluation results is not a necessary goal of evaluation. Evaluations are designed to address specific decisions within specific contexts relevant to the program at hand. Efforts to develop results that generalize to other settings, at times, may compromise the relevance of the evaluation to the particular setting in which the program operates. The importance of generalizable findings will vary with the particular evaluation questions under study.

When program sponsors or clients seek causal and/or generalizable claims about program effects, they should undertake research studies to address these issues. They should not compromise the evaluations process by loading it with an inappropriate agenda.

Chapter 5

The design alternatives available to an evaluator can be organized into three categories: experimental/quasi-experimental, correlational, and ethnographic/qualitative. The essential requirements for the true experiment are randomized assignment of people to treatment and the use of a control group. Quasi-experimental designs do not satisfy the strict requirement of an experiment; they generally do not protect against all threats to internal validity. However, they are frequently the more practical choice in a setting where a true experiment cannot be conducted. While not "true" experiments, quasi-experimental designs can provide strong bases for claims of cause and effect.

Correlational designs are based on the identification of consistent covariation among measured variables. Correlation cannot be interpreted as causality. However, when interpreted within a conceptual framework, it may identify a pattern of relationships useful as a basis for program decisions. Correlational designs are frequently used with intact groups, when random assignment of people to treatment is not possible. This is frequently the case in educational and human service programs, when participation of the target group is voluntary or based on criteria other than random selection.

Ethnographic/qualitative designs emphasize the identification of the meaning that program participants assign to the activities and events in which they are engaged. It emphasizes the development rather than the testing of the theory. It should be noted, however, that qualitative data collection techniques (as opposed to an ethnographic design), such as interviews, can be used with a variety of the design alternatives discussed above.

Weiss (1972) provides a useful discussion of specific design alternatives available to an evaluator.* These include:

- experimental
- quasi-experimental
- time-series
- multiple time-series
- after only
- after only with comparison group
- non-equivalent control group
- patched-up design
- one project, before and after

---

* A detailed discussion of design alternatives is beyond the scope of this paper. However, readers are referred to Weiss, C., Evaluation Research, Prentice Hall: Englewood Cliffs, N.J., 1972, and Babbie E., The Practice of Social Research (fourth edition), Belmont California: Wadsworth Publishing Co., 1986, for general discussions of design alternatives.

Each design offers strengths and corresponding weaknesses. These and other designs appropriate for use in evaluation studies are discussed more fully by other authors.

Fitting the design to the purpose of the study is the basic issue in choosing a design. Given that basis, six issues should be considered when choosing a design:

- rigor
- internal validity
- external validity
- relevance
- appropriateness
- feasibility

Rigor is concerned with the extent to which the study followed proper procedures and adhered to the conventions of systematic inquiry. Internal validity is concerned with the extent to which the treatment, rather than some other plausible rival interpretation, caused the observed effect. External validity refers to the extent to which the findings are applicable to groups and individuals beyond those included in the study. Relevance concerns the extent to which the design fits the purpose of the study and yields results that address the evaluation questions. Appropriateness is concerned with the extent to which the design fits the social and economic context of the setting. Feasibility pertains to whether the proposed evaluation activities can reasonably be implemented in the setting.

The selection of a design often involves trade-offs among these attributes. For example, efforts to increase the rigor of a design may involve increasing experimenter control over the conditions of the treatment--which may, at the same time, reduce generalizability of the results. Likewise, as increased generalizability is sought by trying to keep a program setting as "life-like" as possible, control over the conditions of treatment are reduced.

G. Selection of Data Collection Procedures and Collection of Data

Evaluation studies draw upon a wide range of data collection procedures, such as:

- questionnaires
- tests
- rating scales
- Q sorts
- counting systems
- observations systems
- participant observation
- interviews
- anecdotal records
- archival records
- simulations
- unobtrusive indicators

Frequently, it is possible to collect data that address the same criteria with several different data collection techniques, just as one technique may collect relevant data on several different criteria. The most successful evaluators are those who have a wide repertoire of skills in data collection and are versatile in their use of many techniques. This section will discuss the need for multiple measures, criteria for instrument selection, and special problems in data collection often encountered in developing countries.

Four reasons support the use of several types of data collection instruments--the use of multiple measures--within a single study:

1. Instruments differ in the particular dimensions of an issue they measure. The use of multiple measures provides a means of triangulating, or viewing the same issue from several dimensions.

2. Instruments are limited by their psychometric properties. The use of multiple measures can help compensate for the limitations on the reliability and validity of any single instrument.

3. Stakeholders differ in what they think is important about a project and in the criteria they use to judge project worth. Multiple measures provide a way of increasing the range of criteria addressed in the evaluation.

4. Key audiences of an evaluation differ in what types of data they believe and in the credibility they assign to different measures. For example, the personal testimony of program participants collected through interviews is most persuasive to some stakeholders, while a more quantitative analysis is preferable to others. Multiple measures respond to the differing beliefs of various audiences about the credibility of data. The use of multiple measures increases the likelihood that the study results will be credible across different stakeholder groups.

Having committed to the use of multiple measures, the evaluator still must choose which particular measures to use. Four criteria guide instrument selection:

1. Validity is the extent to which an instrument measures what it is suppose to measure. The various types of validity indicate how well a measure is capable of achieving certain aims. Construct validity is the degree to which scores on a measure permit inference about underlying traits. Content validity refers to how well a measure captures objectives of a program. Criterion related validity, of which there are two types-- predictive and concurrent--refers to how well scores correlate to an external criterion. Predictive validity is when the

criterion is actual performance of the task being measured. Concurrent validity is when the criterion is a score on the another test that purports to measure the same or a similar outcome.

2. Reliability is the extent to which an instrument yields a consistent measurement over time. Reliability can be lowered by such factors as ambiguous instructions, poorly worded items, variation in the measurement environment, etc. Different methods of determining reliability coefficients account for different sources of error. Internal consistency reliability (Kuder Richardson 20, Kuder Richardson 21, split half) is an estimate of the homogeneity of items on a test. Test-retest reliability is a measure of consistency of a measure over time. Parallel forms reliability indicates the correlation between two forms of the same measure.

3. Appropriateness refers to the extent to which measuring instruments are relevant and acceptable for use in the intended setting. It is concerned with issues of both logistics and cultural sensitivity. For example, written questionnaires generally are not appropriate for illiterate respondents, pre-school children, or vision-impaired respondents.

4. Availability refers to whether valid, reliable, and appropriate tests can be secured within the available time or circumstances. Further, if scoring or interpretation cannot be assured within a time frame that meets the needs of the study, there is little point in using the instrument.

One of the greatest threats to the value of evaluations conducted in developing countries is poor data quality. Poor data quality stems from three primary sources:

1. Appropriate instruments may not exist, so new instruments (tests, questionnaires, interview protocols, etc.) have to be designed. Under the pressures of limited time and resources, the instruments may be developed without sufficient consideration of psychometric qualities of the measure. The need to build in time for instrument development and validation should be a consideration of budgeting and scheduling for any study.

2. Insufficient attention is given to procedures for translating data sources (reports, documents) into English and measurement instruments into the local language. The translators used by the evaluation team may not be completely fluent in English or may have trouble translating technical terminology from one language to another. Claims the evaluator makes about the program may be due to misunderstandings and errors of translation.

3. Evaluators may fail to understand the cultural sensitivities of potential respondents and/or the logistical realities of the setting and use inappropriate data collection techniques.

The seriousness of this last problem in international settings cannot be minimized. The standard repertoire of data collection methods run up against serious problems, but at a level of subtlety that often escapes the less experienced evaluator.

By the time a study is designed, the data collection instruments developed, and the sample drawn, the evaluator has a heavy investment in the study and may be reluctant to acknowledge the potential threats to data quality due to cultural bias, insensitivity of the instruments, or reactivity of the data collection methods. Or, given the investment in time and resources that has been made on the study, the evaluator may decide that the bias is "acceptable," often without really understanding what or how much bias is being accepted.

Much of the problem in data collection stems from three sources. First, the infrastructure for collecting the data may not exist. Mailed surveys cannot work in countries in which the postal system is nonexistent or unreliable. Telephone interviews presuppose that the desired sample population has telephones, that the telephone system works, and that the evaluator can secure the necessary phone numbers. Questionnaires have limited use in countries in which large portions of the respondents are illiterate. At the school level, regular record keeping procedures often are not established and forms for recording data, such as attendance, absenteeism, and participant achievement, do not exist or are not widely used.

Second, attitudes that would support individual participation in a data collection exercise may not have been developed. Respondents hesitate or refuse to complete questionnaires or participate in interviews; or, if they do participate, they do not provide candid answers. They may not offer critical or negative opinions for fear that their supervisors or other authorities will find out. Or, respondents may fail to keep necessary records because of concern that the data will somehow be used against them--to embarrass them if their students did not do well, to be used as the basis for cutting resources to their part of the program, or to cause them to have to attend an inservice training program that would threaten their leisure time or their ability to earn a second income.

Third, the data collection methods used may be reactive. The interviews and questionnaires may ask people to express opinions and attitudes that the respondents think inappropriate to express--such as criticism of a teacher or a program administrator. Interviewers may fail to understand cultural preferences to answer questions slowly and indirectly. Thus, the interviewers may terminate the interview before the issues under discussion have been explored adequately.

In some evaluation settings, no acceptable solution to problems of data quality may exist. The most desirable and relevant data collection strategies may be inappropriate or unworkable in specific cultural settings. This may shape the nature of the evaluation questions that can be addressed and/or the confidence clients can have in the study results that are presented. Three principles should guide the evaluator faced with decisions about data collection procedures:

1. Evaluators must be explicit about the quality of the data that supports their conclusions. All evaluation reports should contain a section on data quality in which those limitations are clearly presented.

2. Data relevant to any particular criterion frequently can be collected through several different data collection techniques. Evaluators in cross-national settings must be able to employ a range of data collection techniques to address the same issue. Evaluators who are ideologically opposed to quantitative or qualitative techniques do little to help improve evaluation practice in less developed countries. Expertise in the use of multiple techniques is a prerequisite for evaluators working in developing country settings.

3. In addition to the ability to substitute one measure for another, the evaluator must seek ways to employ multiple measures. Using multiple measures which allow for a triangulation of data collected through several techniques compensates for the limitations of any one date collection technique.

H. Data Analysis

Two major issues confront evaluators in LDC settings at the point of data analysis. First, the local capacity to analyze data in developing countries is often scarce. Second, key audiences of the evaluation frequently have little experience or training in understanding evaluation results.

Low capacity to analyze data occurs for a variety of reasons. For example, there is often a lack of the needed computer equipment or software desired to conduct the analysis. When such equipment is available, it frequently is subject to heavy demand, and the evaluator (often an outsider) encounters competition for computer time from local staff who have a higher priority to use the equipment. Also, in some countries, access to the necessary computer capacity is restricted for reasons of internal (national) security. Complex clearances are needed before access is allowed. In addition, there is often a lack of personnel trained in data entry and statistical analysis. This lack of trained personnel can affect even those evaluators who intend to do their own analysis.

Limited capacity to analyze data places constraints on the types of data that appropriately can be collected and used. It may limit the range of evaluation questions that the evaluator believes can be pursued. As a result, evaluations may be designed

with greater attention to the evaluator's ability to analyze data easily rather than to the importance of the questions being evaluated. Such occurrences fuel the criticism that evaluation results often are irrelevant to the needs of educational planners. The other risk, of course, is that data are collected on key evaluation questions but can not be rendered into a useful form. Thus, they also end up having no impact on planning decisions. It is important, then, for the evaluator to carefully plan how data will be analyzed early in the evaluation design stage. The potential trade-offs between the importance of some evaluation questions and the ability to analyze the relevant data must be addressed before resources are committed to data collection.

Key audiences of the evaluation frequently have little experience or training in understanding quantitative results. Many local government personnel may have moved to high levels of responsibility through their administrative competence, but still lack formal training in statistics, measurement, research design, and policy analysis. Concepts like random sampling and generalizability are hard for some policymakers to understand. Many do not understand inferential statistics--concepts of correlation, multivariate analysis and significant differences may have little meaning. Many clients lack background in key analytic concepts--for example, the difference between unit versus cycle costs of schooling. The claim that it takes an average of 12.8 student-years of schooling to produce one primary school graduate leaves some clients wanting to meet the student who attended primary school for 12.8 years.

One approach is to simplify the analysis to a level which is more easily understandable, e.g., the use of frequency distributions rather than correlations to portray a finding. Simplification of evaluation results should occur to the maximum extent that is reasonable. However, some of the more important questions about education and development cannot be reduced to simple statistics and, in some cases, oversimplification may do more damage than no study at all. Questions about the correlates of program participants' achievement, retention, etc., are necessarily multivariate in nature.

An alternative approach is to build in opportunities to discuss evaluation results with the client and key audiences. This requires that results either be available while the evaluator is in the field or that follow-up opportunities for the evaluator to meet with key groups be built into the evaluation budget and schedule.

I. Interpretation and Reporting of Evaluation Results

In reporting evaluation results, the evaluator encounters a series of issues that are both political and technical. The value of careful planning, thoughtful formulation of the evaluation questions, and rigorous data collection are lost if the evaluator is unable to present the results in a meaningful and credible manner. Among the issues the evaluator must address at this point in the study include:

- How can results be presented so that audiences understand them?
- How can they be reported so that they reach the key audiences?
- How far should the evaluator go in formulating recommendations?
- What is the evaluator's responsibility in presenting the caveats and limitations of the study?
- How should negative results be reported?
- What is the evaluator's responsibility when evaluation results are misinterpreted, distorted, or used for personal or political advantage?

The concern for how results will be reported must start long before the data are collected. Effective reporting requires that:

1. The study be organized around relevant evaluation questions;
2. The planning and logistics provide for the completion of the study and presentation of result within a time frame which allows the results to be available when needed;
3. The key groups with an interest in the study be clearly identified and that the evaluator, client, and sponsor are clear about who is to receive the evaluation report and who will take responsibility for sharing it with other stakeholders;
4. The evaluator understands the abilities of the primary audiences to comprehend technical results so that findings can be presented in a manner best suited to those audiences; and
5. The role of the evaluator in reporting results versus making recommendations be clearly understood.

These issues are ones that should have been clarified early in the working relationships with the client and key audiences. At the time results are ready to be reported, the evaluator faces a series of additional decisions. Four of particular relevance to program evaluation in developing countries will be discussed:

1. Components of an evaluation report;
2. Reporting styles;

3. Reporting formats; and

4. The evaluator's role in making recommendations.

Components of an evaluation report. Figure Three offers a suggested outline of an evaluation report. Four characteristics are of special note. The report should begin with an executive summary which clearly presents the key findings of the study. Secondly, the report should describe the program being evaluated and the economic, cultural, social, and historical context in which it operated. One use of evaluation results is to identify particularly successful program components for possible replication elsewhere. However, knowing a program component was successful is of little help to future program planners if they do not know the setting or conditions in which it operated. Knowing that the ministry officials, participants, and sponsor all regarded a program as highly successful is of little help to those who may wish to replicate it if it is unclear what activities actually constituted the program.

Third, the criteria, standards, and indicators employed in the evaluation should be stated clearly. It should be evident to the reader what issues were considered and what types of data were collected to address those issues. Fourth, the results should be presented and interpreted. It is the evaluator's responsibility not only to report the evaluation results, but to say what they mean in terminology that the client and primary audiences will understand. It is of little help to a client to know that the R2 between program participation and achievement was .39, if they do not know what R2 represents.

One of the most challenging aspects of conducting program evaluations in developing countries is coping with the lack of data and their variable quality. Decisions and policies frequently are based on data that are incomplete or flawed, but which is, nonetheless, the only (and consequently, the best) data available. An essential element in all evaluation reports is a section on data quality. It should describe the source of the data, limitations and weaknesses encountered in collecting and analyzing it, and caveats that should be observed in interpreting the results. The evaluator holds special responsibility for completing this section since the evaluator frequently is the only one aware of the extent of problems. At minimum, this section should report the extent of missing data, problems in operationalizing constructs used in the study, validity and reliability limitations of the instruments, other assumptions and judgments made during the study, and implications of these limitations for interpreting the results.

It is not uncommon for the client to resist inclusion of this section for two reasons. First, if results favor the position of the client, the client may not want to undercut the credibility of the conclusions. Second, this section tends to appear excessively technical. Clients are not always sure of how to evaluate the potential impact of the data quality problems on the results of the study. They resist inclusion of a discussion

that sounds equivocal. It is important, then, that the evaluator serve as the proponent for inclusion of this section in the final report.

FIGURE THREE

Suggested Outline for an Evaluation Report

Section I: Summary

- What was evaluated?
- Why was the evaluation conducted?
- What are the major findings and recommendations of the evaluation?

Also, if space permits,

- Were there decisions to be made on the basis of the evaluation? If so, what decisions?
- To what audiences is the evaluation report addressed?
- Who else might find it interesting or important?
- What were the major constraints, if any, under which the evaluation had to be carried out?

Section II: Background Information Concerning the Program

- Origin of the program.
- Goals of the program.
- Characteristics of participants.
- Characteristics of program materials, activities, and administrative arrangements.
- Who is directing/running programs?

Chapter 5

Section III: Description of the Evaluation Study

- Purposes of the evaluation

  - Who requested?
  - Who are audiences?
  - What kind of information did the audiences require?
  - Etc.

- Evaluation design(s).

- Outcome measures

  - What outcomes were measured for the evaluation?
  - What data were collected?
  - Etc.

- Data collection procedures

  - What instruments were used?
  - What was data collection schedule?
  - From whom was data collected?

Section IV: Results

- Present the results of the various measurements described in Section III.

- Interpret the meaning of these results.

Section V: Discussion of Results

- Interpretation of each result occurs in Section IV. However, if the evaluation is complicated, a special section discussing the results makes the report clearer. Results should be discussed with particular reference to the purposes of the evaluation listed in Section III.

- Two major issues to be addressed here:

  1. How certain is it that the program caused the results?
  2. How good were the results of the program?

Section VI: Costs and Benefits

Section VII: Conclusion and Recommendations

- Conclusions.
- Recommendations regarding the program.
- Recommendations concerning subsequent evaluations of the program.

Source: L.L. Morris and C.T. Fitz - Gibbon, How to Present an Evaluation Report, Beverly Hills: Sage, 1978

Reporting Styles. One of the key issues in reporting results is how formally the findings should be presented. Should findings be shared with the client as they emerge or should silence be maintained until the conclusion of the study? There are tradeoffs in either direction, and practice will vary with both the characteristics of the client and the role the evaluation is playing in the program. Typically, formative evaluation results may be shared through conversations and memos to the client, while summative evaluation is almost always presented in a more formal manner. However, the growing literature on the utilization of evaluation results emphasizes the importance of the rapport and informal communication between the evaluator and the users of the evaluation, regardless of the particular role of evaluation in the program (Alkin and Dailluk, 1979). Evaluation results are most likely to be used if:

1. the evaluation addresses issues important to the user, and the evaluator presents the results in a form preferred and understandable to them;
2. the evaluator encourages user involvement in the design of the study;
3. the evaluator explicitly tries to facilitate and encourage the use of the results;
4. the evaluator is credible to the client and key audiences; and
5. a dialogue (as opposed to one-way communications) regarding the evaluation findings takes place between the evaluator and the user.

Evaluators of international technical assistance programs encounter several special problems. First, evaluation studies are seldom completed in the field. Limitations of time, the extra costs of field work, and the unavailability of the equipment needed for data analysis often require the evaluator to complete the final report after leaving the country. This pattern seriously limits the dialogue between the evaluator, client, and

stakeholder groups. It encourages one-way communications in which the final report is returned to the field without opportunity for the client and stakeholders to discuss the results with the evaluators.

A second problem is that sharing intermediate evaluation results may give rise to a series of lobbying activities by interest groups to encourage the evaluator to suppress or alter the findings. Consequently, the client may not want information to be shared. This occurs particularly when the client is concerned that the premature disclosure of results will foreclose later options for responding to or ignoring the evaluation.

A third problem is one of language and logistics. Key audiences of a report may not be sufficiently fluent in the language of the evaluation or fully understand the arguments, evidence, and logic that support the results. Consequently, the report may not receive the attention that the results and implications of the study deserve. A fourth problem is posed by disagreements among the team members. When a minority of the team disagrees with the majority, or a majority disagree with the team leadership, decisions regarding which results to report and the manner in which they will be reported can be difficult to resolve. Disagreements among team members, when they are detected by stakeholders of the evaluation, may be exploited so as to discredit the evaluation effort or to minimize the evaluation results.

Whenever possible, resources should be provided for the evaluator to complete a substantial draft of the report prior to leaving the country and for the client and key audiences to read and discuss it with the evaluator. Where this is not feasible, arrangements should be made for the evaluator to return to the country after the report is in final draft form to discuss it with the key groups involved. This provides an opportunity for errors of fact or omission to be corrected, for differences in interpretation to be debated, for the client and other audiences to understand the evidence, results, and recommendations more fully, and for these groups to feel some ownership of the study.

Reporting Formats. Among the most important issues in evaluation utilization are the ability of the client and audiences to understand what was done in the evaluation. Just as audiences differ in the criteria they hold important, the standards they employ, and the types of data they trust, they frequently differ in their ability to understand technical concepts and terminology used in reporting evaluation results. Reports may have to be "tiered," so that results are reported at several levels of audience interest and technical sophistication. Alternatively, evaluators may need to develop several reports, tailored to the interests and information needs of key groups.

The extent to which the study makes recommendations based on the findings will vary, based on initial agreements with the client as to how the report will be used. In most situations, it is advisable for the evaluator to develop recommendations because it is the evaluator who best knows the data, has the technical skills to move from technical interpretation to policy suggestions, and often is the most impartial person

within the field setting. Nonetheless, the evaluator must work closely with the client in developing the recommendations so that the client can understand more fully the evidence, logic, and rationale that supports the recommendations and so that the recommendations are presented in a language and format that most encourages adoption.

CHAPTER SIX

TECHNICAL ISSUES OF EVALUATION IN THE DEVELOPMENT CONTEXT

I. INTRODUCTION

Sometimes the technical decisions made concerning how evaluation data will be collected or analyzed inadvertently shape and/or change the evaluation question being investigated or the meaning that can be assigned to the evaluation results. Only too late does the evaluator realize that the evaluation results are not appropriate to the policy issue. Also, the evaluator may not detect the impact of subtle technical decisions on the evaluation and may report erroneous results without being aware of the implicit biases resulting from the earlier technical decisions.

Four design issues are discussed in this section:

1. The appropriate unit of analysis and level of aggregation in evaluation studies.
2. Dilemmas in the measurement of change.
3. The inadequacies of testing.
4. Translation procedures for the cross-cultural use of measuring instruments.

These design issues will be discussed in the context of structuring the evaluation process to apply to international technical assistance programs.

II. ISSUES IN LEVEL OF AGGREGATION, UNIT OF ANALYSIS, AND CROSS-LEVEL INFERENCE

What is the appropriate unit in which to analyze evaluation data? Should data be analyzed at the student, classroom, school, or district level? These questions arepractical ones because variables may take on different meanings at different levels of aggregation. Hence, the level of aggregation at which the data is analyzed may alter the results and interpretations that might follow from the analysis.*

Comprehensive discussions of multilevel data analysis are provided by L. Burstein, "The Role and Levels of Analysis in the Specification of Educational Efforts," Chapter 3 in R. Dreeben and J.A. Thomas, The Analysis of Educational Productivity, Volume I: Issues in Microanalysis, Cambrige, Mass.: Ballinger, 1980, and L. Burstein, "The Analysis of Multilevel Data in Educational Research and Evaluation," Chapter 4 in Review of Research in Education, Vol. 8, 1988. A recent application of contextual analysis is provided by K. Eksterowicz, Contextual Effects and Relationships in Student Ratings of Secondary Instruction in New York State, Unpublished Doctoral Dissertation, School of Education, Albany, N.Y.: State University of New York at Albany, 1985.

This concern with the unit of analysis leads to a second issue-- that of cross-level inference. Most educational programs occur within some group context. Students are located within classrooms, within schools, within districts, etc. Educational influences on the student are shaped by the groups to which the student belongs. An important question in evaluating program impacts is often the extent to which behaviors of the individual (such as learning) are affected by the classroom, school, and district in which the individual is located.

These analytic issues of compositional context are well represented in the statistics and social science literature; yet, they are frequently overlooked in practice. It is not uncommon for studies of student achievement to employ large regression analyses in which the appropriateness of the unit of analysis is given little attention. The following discussion will address three issues:

1. unit of analysis,
2. cross-level inference, and
3. contextual effects.

Unit of Analysis. A measured variable can take on different meanings at different levels of analysis. For example, family income and father's employment, at an individual level of analysis, may both serve more appropriately as proxies for socioeconomic status of the family. When aggregated across the community, however, they serve as indicators of community resources. A high level of community resources reflects, indirectly, special benefits that may be available to all students in the school, regardless of the particular wealth of a student's family. Similarly, the amount of encouragement a student receives from parents, at another level of analysis, may be reflective of a community orientation toward education. What might be measured at one level as an individual's attitude may, at another level, represent the environmental press of the school.

In conducting an educational evaluation, it is important that educational data be collected and analyzed at the appropriate level of aggregation (i.e., unit of analysis). The decision as to the appropriate unit, however, is not always easily made. The evaluator encounters two issues: (1) single-level analysis is seldom appropriate for multi-level educational data and (2) the appropriate unit of analysis depends on the nature of the questions of policy or practice under investigation and the substantive conceptual model being used to guide the inquiry. The specification of multi-level effects requires strong substantive theories to guide the selection of variables and the level at which they should be analyzed. Such models are still at an early stage of development.

Cross-Level Inference. Cross-level inference involves making inferences about relationships at one level from relationships found in data at a different level. When

going from individual to group inferences, a frequent issue is statistical dependency. When going from group to individual inference, the issues raised are of contextual effects. In both cases, there is significant danger that a compositional fallacy may ensue.

The issue of statistical dependency arises when analyses are based on individual data, but individuals experienced the treatment within intact groups (such as classrooms or schools). When intact groups are assigned to instructional treatments, the individuals in those treatments should not be considered as independent units. The lack of independence among observations violates the assumptions of some common statistical tests (e.g., regression).

The practical implications of these issues in terms of evaluation design are enormous. For example, suppose that as part of an evaluation of the efficiency of an inservice teacher training program, secondary students' ratings of their teachers' effectiveness were correlated with student achievement. Assume the rating and achievement data were collected for 30 students in each of ten classes and that the analysis was conducted using the student (N=300) ratings to predict student achievement.

The individual student exists as the unit of analysis. The problem, however, is that observations across students in the same class cannot be independent. That is, those students within a class shared an experience that affected their individual performance, and this experience was not shared by students in other classes. They shared the teacher, a particular classroom environment, a peer group, and a unique school setting--all of these different from that experienced by students in the other nine classes. Many statisticians would argue in such a case that the classroom is the more appropriate unit of analysis. This would entail using the class mean on the rating instrument to predict mean classroom achievement score.

Shifting to the classroom as the unit of analysis poses different, but equally severe problems. Instead of conducting the study over 300 observations (students), the analysis is now conducted over ten observations (classrooms). Use of more general organizational units of analysis reduces the number of observations available as input to the analysis. The drop in the available degrees of freedom often precludes the use of the desired statistic--particularly the multivariate statistic--because the degrees of freedom for the analysis drop below the minimum acceptable sample size to support a stable statistical test. The obvious solution to this is to increase the sample size by including more classrooms in the study; however, this has financial and social implications that may be unacceptable.

Contextual Effects. Contextual effects refer to the effects of groups on individuals. A contextual effect for ability, for example, is said to occur when group mean ability is related to individual outcomes after controlling for individual ability. The use of group level variables to predict individual level outcomes is common in education. Burstein

(1980) points out that properties of teachers, classrooms, and schools are at a macro-level with respect to students within those classrooms and schools. For example, evaluations of educational programs might want to examine the quality of school facilities and the adequacey of teacher training on student achievement. Indeed, much of the research on educational effects involves prediction of individual level outcomes (achievement) from group level predictor variables, concentrating on the school, teacher, and classroom effects on student achievement.

III. PROBLEMS IN THE MEASUREMENT OF CHANGE

Educational programs are conducted to bring about a desired change--in the way people do their jobs, in what they know or the skills they can apply, and in their attitudes, beliefs, or outlooks. One of the most intuitively appealing notions to many program sponsors and administrators, feeling the press for accountability, is to demonstrate participants' gain as a consequence of program participation. Consequently, many evaluations have sought to demonstrate the change over time in the particular abilities, skills, or attitudes of participants that might be attributable to the program. Their effort involves three conceptually distinct activities:

1. measuring attainment,

2. measuring achievement, and

3. developing an attributional claim, e.g., that the achievement was related to or caused by participation in the program.

Change, however, is an ambiguous concept. There are two areas of confusion. One is conceptual--the concept of attainment and achievement are misunderstood and change, gain, and difference scores are confused. The second is statistical--serious technical problems are encountered in efforts to measure value added as a change concept.

Attainment refers to performance at one point in time. Achievement refers to the change in performance over time, the change in attainment from one point to another. Attribution is a function of the evaluation design that was used and is not a measurement issue per se. If program participants are randomly assigned to treatments, an evaluation may only examine differences in attainment among groups receiving different treatments to form a conclusion about the effectiveness of an educational program. The issues of achievement need not arise.

Frequently, however, in the study of educational efficiency the issue of interest is achievement. The focus on achievement occurs when the evaluation question is concerned with: (a) rate or amount of gain as a function of treatment, (b) selection of individuals on the basis of gain scores (for example, identifying fast learners for special

opportunities), and (c) in non-random designs when it cannot be assumed that groups were equal in their initial attainment.

Change, or difference, scores refer to alteration in behavior (e.g., learning, performance, etc.) between two points in time. A gain score is calculated by subtracting the pretest score from the posttest score. It includes (and inherently confounds) score variation due both to change and to measurement error.

The statistical estimation of change poses a serious problem in measurement. Substantial literature supports the theory that gain scores are rarely useful, no matter how they might be refined. Cronbach and Furby (1970), in their classic article about the measurement of change, conclude that investigators who ask questions regarding change ordinarily would be better advised to frame their questions in other ways.

Five approaches to measuring change are discussed in the evaluation literature: crude gain scores, change in group means, residualized gain, residualized true score estimates, and analysis of covariance.*

Crude Gain Scores. Raw gain or difference scores are computed by subtracting pretest scores from posttest scores. Use of these scores leads to false conclusions because such scores are systematically related to any random error of measurement. The gain score picks up any unreliability in the pre- and post-measures. Stake (1971) offers a dramatic example of what this means in practice. Suppose pre- and posttesting was conducted using parallel forms (Test A and Test B, respectively) each of which had a reliability of .84. The correlation between Test A and Test B was +.81. The reliability of the gain score would be +.16, an unacceptably low level. Due to this unreliability, gain scores can imply learning that did not occur or fail to detect real improvements that did occur. Take Stake's example again. Suppose on the test described above, the raw score standard deviation and grade equivalent standard deviation were 9.5 and 2.7 years respectively. On average, using a 95% confidence interval:

- Individual student's raw score would be in error by 2.5 items,
- The student's grade equivalent score would be in error by .72 years, and,
- The student's grade equivalent gain score would be in error by 1.01 years.

For a more technical discussion of measuring change, see L.S. Cronbach and L. Furby, "How Should We Measure Change: Or Should We?," Psychological Bulletin, 74, 1, 1970, 68-80. See also, J. Stanley, "General and Special Formulas for Reliability of Difference," Journal of Educational Measurement, 4, 4, 1967, 249-252. A discussion of measuring change in studies that employ multiple intercorrelated dependent variables is presented by C.W. Harris, "Some Problems in the Description of Change," Educational and Psychological Measurement, 22, 2, 1962, 303-319.

Stake describes the consequences of this level of error:

- Assume that students are allowed to exit the program when their improvement is one grade equivalent or more.
- Assume also that three students are tested with a parallel form immediately after the pretest.
- The chances are better than 50-50 that one of the three students--entirely due to errors of measurement--will gain a year or more and appear ready to graduate the same day they were to begin the program of study.

Change in Group Means. Change, or difference scores, can be formed by subtracting pretest from posttest means for a group, such as a class or a particular school. This approach is an improvement over the use of crude gain scores due to the stability gained with measures of central tendency. Nonetheless, the unreliability of the difference reamins a problem.

Residualized Gain. A gain is residualized by regressing the pretest on the post-test score. This separates the post-test score into variance "explained" by pretest performance and residual, or unexplained variance. The residual represents both the change in performance due to the program and measurement error. The portion of the post-test information that is clearly predictable from the pretest has been partialled out. Cronbach and Furby (1970) point out that the residualized score is not a "corrected" measure of gain, since the portion of variance discarded (e.g., explained) include some genuine and important changes in the person. The residualized gain serves primarily to single out indivuduals who change either more or less than expected.

Residualized True Score Estimates. In classical measurement theory, the score a person receives on a measure ("observed score") consists of two parts---the person's true score (e.g., their real attainment on the construct being measured) and error. Error can be due to many things---unreliability in the measure, variability in the data collection situation, or aspects of performance not captured by the particular measure used. The residualized true score approach to measuring change attempts to improve on the residualized gain by estimating the true scores, both pre and post, and computing gain over only the true score components. While conceptually appealing, use of this method does not yield meaningfully different estimates than the direct computation of residualized gain.

Analysis of Covariance. In this approach, post-test scores are used as the dependent variable while the pretest scores are used as a covariate. For example, assume an evaluation study is investigating the impact of different levels of program participation on level of skill acquisition. The skill attainment measure used as the pretest would

serve as covariate, and the initial level of skill attainment would be controlled. The adjusted dependent measure is a sort of residual gain; however, the procedure does not involve calculating residual gain scores for individuals.

When treatment groups are not formed at random, which is often the case in educational programs where it is common to use intact groups, the covariance approach may mask other problems. Specifically, if the treatment groups differed systematically at the start of a program on any relevant characteristic other than the covariate, the covariate will not remove that confounding factor.

As the above discussion has indicated, the measurement of change is fraught with problems, both conceptual and statistical. The literature supports four generalizations about the measurement of change.

1. All methods of measuring gain are unreliable at the individual level of analysis.

2. Evaluators are well advised to formulate the evaluation question and subsequent evaluation design to examine attainment rather than achievement.

3. The stability of measures of central tendency improve the measurement of change at group levels.

4. The measurement of change in student performance, pre- and post-program, does not constitute evidence that the program was associated with or caused the achievement. The issue of attribution is an issue of evaluation design.

The reasons for wanting to measure achievement are intuitive and appealing--change in achievement and performance often are regarded as the most important effects of a program. The reasons for not measuring change tend to be complex and technical--hard to explain to clients with little background in statistics. Situations arise in which the evaluator does not have a meaningful choice. In those situations, the recommended statistical estimate of change is the measured gain adjusted for entering ability. A difference score is computed as post-test score minus pretest score. Then, the pretest score is regressed on the difference score to remove the variance in the difference score that might be explained by entering student ability. This procedure yields a gain score adjusted for entering ability.

IV. THE LIMITATIONS OF TESTING

Testing is one of the most popular and pervasive data collection techniques in evaluation. It enjoys vast credibility around the world. Four reasons contribute to this. First, tests appear to be highly objective. This is particularly important in developing countries where basing personnel selection decisions on tests is seen as a meritocratic

means of moving beyond political, tribal, or family lines of influence. Entry to educational, training, and career opportunities based on test performance provides a sense of fairness and impartiality to the selection process.

Second, tests are comparative. Developing countries face extraordinary social demand for education and training activities that provide participants with access to opportunities in the modern sector of the economy. Governments are faced with a legitimate and important need to select those who will move into the education, training, and employment opportunities that emerge. Tests provide a basis for that selection. Tests are used to increase efficiency because the "best" students or candidates are combined with available scarce resources.

Third, national testing systems support the centralization of authority. Testing systems are a means of controlling people, resource allocation, and the social order. As governments in many developing nations attempt to secure, consolidate, and extend their influence, centrally controlled testing systems provide one reasonably efficient means of controlling the distribution of opportunities and social rewards.

Finally, many countries have a long tradition of student testing inherited from their colonial era. Testing systems predate other evaluation techniques currently advocated for use in developing countries. Testing was introduced by the colonial administrations to screen applicants for both education and employment opportunities, as well as to evaluate the quality of the schools and colleges.

Testing is an important tool and can play a legitimate role in effective evaluation. However, it is widely misused, largely because the nature of the claims that can be made on the basis of a test score are poorly understood. The appropriate use of tests requires clarity on nine issues:*

1. Testing is not evaluating. Testing is a form of measurement. Measurement allows us to quantify an entity or construct, such as achievement or aptitude. The meaning of that measurement is not inherent in the quantification. Rather, its meaning is something that people assign. Evaluation has not occurred until the measurement has been interpreted and a judgment about the meaning of that

This section draws particularly on the work of E.F. Kelly, "The Role of Testing in American School Reform," in National Educational Reform and New York State: A Report Card, Rockefeller Institute Conference Proceedings, Albany, N.Y.: State University of New York at Albany, 1985. Kelly offers an insightful policy analysis of the use of testing in schools that raises issues of relevance far beyond American classrooms. See also E.F. Kelly, "Horse Races, time trials, and evaluation designs: Implications for future evaluations of the Improved Efficiency of Learning Project," School of Education, Albany, N.Y.: State University of New York at Albany, 1984.

score has been rendered. Tests do not make decisions, people do, and they do so only within a larger conceptual structure in which the test measure is only a minor item.

2. Large scale standardized tests, the type most often used in developing countries, do not tell us what students know. Two students who get the same score do not necessarily know the same things. They did not necessarily answer the same items correctly. Students' scores do not allow us to specify what the student knows, but rather, allows us to place the student-- in a distribution of other students who also took the test--in terms of the relative probability of what they

3. Tests do not measure learning directly. Test items are proxies for the behaviors and knowledge one wished to assess. Further, items on a test are only a sample of the universe of possible items that might have been asked. The claim, based on a test score, of what participants have learned is actually an inference to that universe of items of which the test is but a sample. Also, the strength of claims about learning rests on the extent to which the test items represent actual behaviors or knowledges.

4. Test scores are not perfect measures of knowledge or achievement. All tests have measurement error and are less than perfect approximations of the construct or knowledge they attempt to assess. Error, in this sense, refers to variation in test scores due to factors other than the person's command on the content. Factors such as ambiguous items, unclear directions, and adverse testing conditions contribute to measurement error. There are well understood procedures for estimating the measurement error in tests. Often, however, these procedures are not employed or, when the extent of measurement error is computed, those responsible for interpreting and using test results do not understand the practical meaning of the measurement error statistic.

5. Often the domain one seeks to test is poorly understood. Test items are a sample of a universe of possible items that represent the construct or the knowledge being tested. If the construct or content domain is not clearly specified, no basis exists for claiming the test items represent the domain.

In many developing countries a common example of an ambiguous content domain is the national curriculum. In recent years, most countries have formulated a national curriculum and are developing national examination systems to monitor student performance on that curriculum. Two problems have occurred. First, the curriculum has been specified as general goal statements, rather than as specific learning objective. No agreement at an operational level exists as to what the goals mean. Second, these goal statements, while disseminated to local schools, are not directly supported by

instructional materials available in the schools or by the existing supervisory systems.

Consider a recent example: As a part of a donor funded curriculum improvement project, an external testing agency was commissioned by the Ministry of Education to develop a standardized test to measure student achievement in areas specified by the national curriculum. Using this test in a post-test only design, students in the treatment did not differ significantly from the comparison groups. The program administrators argued that the new instructional program was an improvement over the instruction offered in non-treatment schools and that the test must have been deficient. It was claimed that the test lacked curricular validity. The program staff developed a second test designed to measure the national curriculum, but which assessed specific knowledge and skills taught in the experimental program. When students from the same treatment and comparison schools were again tested, a statistically significant difference between the treatment and comparison schools was observed. The program administrators accepted this as evidence of the success of the new instructional program and used these results to argue for further funding. The example illustrates three issues:

- Both tests claimed to operationalize the national curriculum. That is, they claimed to be different samples of items drawn from the same universe.

- It appeared that the two tests were measuring different constructs and/or knowledge. Re-analysis of the data found that, within the same grade level, English and math test scores correlated more highly with each other than the English test developed by the external group correlated with the English test developed by the internal program team. The same was true for the math tests (Kelly, 1984).

- The choice of which test to believe would lead to very different decisions about the future of the project. Results of the two tests support different policy alternatives.

The situation described in this example, in large part, is due to the ambiguity and lack of specification of the content domain being tested.

6. <u>There are insufficient bases for establishing meaningful standards</u>. Standards address the questions of how much is enough and what level of achievement will be regarded as adequate. In education there are multiple standards and different groups frequently disagree on the most appropriate standards to use. Moreover, the same group may use different standards at different times with respect to the same program. A level of achievement held to be sufficient by one standard

(i.e., student performance in the same school a year earlier) may be judged inadequate by another (i.e., the performance of students in another region).

A recent examnle illustrates the issue. In one country, results of the national secondary school leavers examination were reviewed annually by a committee (appointed by the central government) which adjusted scores upward to help reduce the political pressure on government personnel that would result if too low a percentage of students passed the test. A high ranking Ministry of Education official disbanded the committee in an effort to increase equity and the credibility of the secondary school diploma. When the committee was disbanded, the pass rate on the examination dropped. Reaction from parents was intense and the MOE re-established the committee. Key stakeholders in this example differed in the standards they held for test performance.

7. Achievement (as measured by tests) may not be the most appropriate criteria on which to judge a program. Testing is widely understood, relatively easy to accomplish, and seemingly central to the intent of many programs. The credibility and relative ease of testing may allow it to overwhelm proper analysis and displace attention from potentially more appropriate and important criteria. These might include the extent to which a program is implemented, the quality of instructional activities, or the time spent on learning tasks.

   Tests often are designed to measure what students are taught. However, as Clark and Voogle (1985) point out, transferability of learning often is more important than mastery of knowledge. The trainees' ability to apply a concept in a new and novel situation is more important than their ability to respond to criterion oriented tests. This preoccupation with testing by evaluators and program administrators often leads them to forget other important processes and outcomes of the programs.

8. Program quality cannot be improved by raising test performance standards. Changing standards may alter the number of persons admitted to or allowed to graduate from a program, but does nothing to change the instructional quality of the program itself. Serious efforts to improve educational quality must address the selection and delivery of curricular content. This may involve improved materials and materials distribution, more efficient use of teachers' time, and improved pedagogy--but altering the cut-off score does nothing to change the quality or efficiency of the program itself.

9. Gain scores should never be used as a basis for program evaluation. As noted above, raw gain scores are computed by subtracting the student's pretest score from the post-test score. Gain scores are highly unreliable and should not be employed for the reasons cited earlier.

Most testing procedures in developing countries are norm referenced. Criterion Referenced Tests (CRTs) provide more accurate information on what trainees know. However, CRT's do not provide the score discrimination often preferred for personnel selection. The predominant use of CRTs has been in vocational/technical training and in training that utilizes curriculum embedded CRTs as teaching tools. However, the increasing use of criterion referenced tests raise equally perplexing issues of practice. The issues of curricular validity, reliability, and standard setting still confront the evaluator using a CRT.

Clark and Voogel (1985) recently have raised several additional questions about the outcomes of criterion referenced testing.* They suggest that testing procedures used in a program may actually contribute to the failure of trainees to transfer their training in desired ways. This happens when a CRT emphasizes procedural outcomes of learning (concrete, practical knowledge of relatively simple routines) in situations in which declarative outcomes of instruction are actually more important. Declarative learning emphasizes far transfer--i.e., the learning of concepts and principles so that they can be generalized to solve new or unique problems.

Clark and Voogel argue that most training technologies implicitly encourage the use of procedural objectives, even when far transfer of skills is desired. This happens when the tests used in a program emphasize criterion oriented items close to the curriculum. These tests have a high level of curricular validity. However, Clark and Voogel suggest that transfer often is more important than mastery.

Consider the example of a management training program for upper and middle level civil servants in a developing country. The goal of training is for the participants to be able to apply what they learn in a variety of new situations. Indeed, the training itself will qualify them for advancement into professional settings and problem situations they have not yet experienced. If progress is measured by performance on tests that encourage recall of procedures and rules in effective management, it may work against the very outcomes most desired--the ability to generalize the principles the participants have learned and to adapt the principles to totally new situations.

Training programs in international settings are particularly susceptible to the problems Clark and Voogel raise. Program administrators are often under pressure to show results of their training within short timeframes to meet the sponsor's demand for accountability. This pressure favors tests that measure near rather than far transfer. Evidence of far transfer is slower to emerge and more difficult to collect than is evidence of short-term procedural learning. Second, there often is a tendency for programs

Clark, R.E. and A. Voogel, "Transfer of Training Principles for Instructional Design," Educational Communications and Technology Journal, 33, 2, 1985, 113-123. The authors present the near versus far transfer dilemma discussed in this paper. They also provide a useful discussion of the differences between behavioral and cognitive approaches to transfer of training and to instructional design.

designed by foreign technical assistance personnel to emphasize procedural learning. This is partly due to the failure of the foreign personnel to understand the culture of their trainees' work settings well enough to draw meaningful examples and illustrations of principles being taught. Locally relevant examples are a part of encouraging far transfer. The training designed by foreign personnel is more likely to teach principles as a series of steps, procedures, or rules with the hope that the trainees themselves can make the transfer to locally relevant situations.

Both norm and criterion referenced testing have an important role as data collection techniques. However, the evaluator carries a major responsibility for insuring that tests are adequately developed, appropriately administered, and correctly interpreted. At minimum, the evaluator must assure: (a) that the test has appropriate levels of validity and reliability, (b) that psychometric properties of a test are computed and reported with the test results, (c) that users of the test information understand the practical implications of the remaining measurement error in making their decisions, and (d) that results are not reported as gain scores. Tests are most appropriate within an evaluation design which emphasizes multiple measures, so that tests are not the sole basis for judgments of program worth.

V. TRANSLATION PROCEDURES FOR THE CROSS-CULTURAL USE OF MEASURING INSTRUMENTS

One aspect of instrument use particularly relevant to international settings concerns the translation procedures employed when a measure developed for use in one country is adopted for use in another country. If proper translation and validation procedures are not employed, evaluation studies suffer from the possibility that the results obtained are biased due to errors in translation and do not depict accurately differences in the people or the variables being measured.*

The borrowing of measuring instruments from another country occurs for four reasons. First, many developing countries have only a short history of empirical research and have few instruments developed specifically for use in the local setting. This is especially true for psychological (e.g., attitudinal) measures developed from a particular construct or theoretical point of view. Secondly, instrument development, correctly done, can be complicated, time consuming, and expensive. Use of instruments with a known history of validity and reliability can reduce the staff time and costs associated with instrument design. Another reason borrowing occurs is that

---

* This section draws heavily from work by D.W. Chapman and J.F. Carter, "Translation Procedures for the Cross-Cultural Use of Measurement Instruments," Educational Evaluation and Policy Analysis, 1, 3, 1979. 71-76. See also Brislin, R.W., W. J. Lonner, and R. M. Thorndike, Cross Cultural Research Methods, New York: John Wiley & Sons, 1973.

many of the persons with evaluation expertise in developing countries received their graduate training abroad and have interests and experience with a certain instrumentation. They sometimes introduce that instrumentation to their country upon their return home. Finally, developing countries frequently are short of personnel trained in psychometrics and instrument development. Even in countries that have social science graduate programs within their own universities, those programs tend to emphasize research design and data analysis skills, rather than the psychometric skills associated with instrument validation.

Figure Four illustrates the type of problem that may occur when insufficient attention is given to verifying the accuracy of the translation. The Classroom Behavior Survey (CBS) was developed and used by Kelly and Chapman (1978) as a measure of American high school teachers' behaviors and classroom characteristics. When it was selected for use in Iran, items were translated to Farsi and the Farsi items were then back-translated into English (Chapman and Kelly, 1981). The discrepancies in the item meanings that were identified in the initial English to Farsi translation highlight the validity problem posed by translation procedures. Had back-translation not been employed, the researchers would have been unaware of the shift in meaning in several of the items.

The most common procedure for verifying the adequacy of a translation is back-translation. The instrument is translated into the second language by one translator, and the resulting translation is then translated back into the original language by a different translator. Discrepancies between the two translations can then be modified and a second back-translation conducted. More than one back-translation should be conducted; two are usually sufficient if the back-translation is consistent with the original version in both iterations. If discrepancies continue to be observed, the translation/back-translation/modification sequence should continue until those discrepancies are removed.

Typically in the translation of an established instrument to a second language, the instrument, as it appears in the original language, is regarded as fixed and invariant. The second language version is altered to reflect working in the original language version accurately. One problem, however, is when a word of concept has no clear equivalent in the other language or when an exact, formal translation may distort the meaning. This may be due to idiomatic usage, language structure, or socially sensitive expressions that are offensive when translated literally. In this case, the evaluator is confronted with the issue of functional versus formal equivalence. If functional equivalence is to be sought, the evaluator may need to tolerate flexibility in the translation. Chapman and Carter (1979) offer an example encountered in translating the Nowicki-Strickland locus of control scale (Nowicki and Strickland, 1973) from English to Farsi. One item read, "If you find a four-leaf clover, do you believe it might bring you luck?" Since the belief in the luck significance of a four-leaf clover is not shared by Iranians, formal equivalence would have resulted in a meaningless translation.

FIGURE FOUR

Examples of Back Translation of Items From the Classroom Behavior Scale

Item 1:	Original:	In this class we never cover all of the materials we are supposed to.
	Back-Translated:	In this class we are not able to keep up with all the assigned readings.
Item 7:	Original:	In this class we cover the same material over and over again.
	Back-Translated:	In this class, similar subject matters become repetitive.
Item 11:	Original:	This class has very little to do with anything that's important to know.
	Back-Translated.	This class deals less with important and valuable subjects.
Item 16:	Original:	This teacher never knows when to stop answering a question.
	Back-Translated:	The teacher of this class does not know how to stop lengthy answers given by students.
Item 33:	Original:	The teacher doesn't involve the students in discussions.
	Back-Translated:	The teacher of this class does not allow the students to participate in class discussions.
Item 63:	Original:	Frequently one or two students monopolize the class discussions.
	Back-Translated:	Most of the time one or two students take a lead in class discussions.

Source: Chapman, D.W. and J.F. Carter, Translation Procedures for the Cross-Cultural Use of Measurement Instruments, Educational Evaluation and Policy Analysis, 1, 3, 1979, 71-76.

However, a similar notion was contained in the Iranian belief that if the seeds of the wild rue plant are thrown into the fire, the resulting puff of smoke will ward off evil. Thus, the eventual translation of the item was, "If wild rue is thrown into the fire, do you think it will keep evil spells away from you?" (Farhad-Motamed, 1979).

The alteration of items during translation in order to preserve meaning requires empirical verification before claims of equivalence can be made. The equivalence of meaning can be checked in three ways.

- One approach is to have bilingual respondents rate both versions for their similarity of meaning.

- A second approach is to investigate the criterion related validity (predictive and/or concurrent) of the second-language measure. However, validation studies can be expensive and time consuming. Time and resources have to be budgeted to accommodate these steps. Further, the estimation of concurrent validity depends on the availability of concurrent measures which, for the reasons described earlier, frequently are not available.

- A third approach is to have bilingual respondents complete both versions (using a counter-balanced order of presentation) and compute the correlation between respondents' scores on the two measures. Correlating overall scores may still mask inconsistencies among items. Hence, a safer approach is to correlate responses across the two forms of the instrument for each item separately.

For any evaluation in which important decisions about people and resources are going to be based (even in part) on data collected through an instrument translated from another language, all three of these techniques should be employed and the results (of the comparison of forms between the two languages) reported in a technical appendix to the final evaluation report.

In evaluations in which instrumentation is being specially developed for use in more than one language (rather than an existing measure being adopted and translated), the evaluator may also employ a procedure, called decentering, in which both the original and second language version of the instrument are subject to modification. This procedure allows for the modification (on either form of the instrument) of concepts that have no clear equivalent in the other language. Decentering would still be conducted in conjunction with back-translation and the empirical verification procedures described above. However, the use of decentering is severely limited in situations in which the original instrument already has a history of use in the original language and is not subject to change without violating the psychometric history of the instrument.

The importance of verifying the accuracy of translation is not limited to measuring instruments but affects all aspects of international work where written materials in one

language are translated for use in another. Of particular concern is the translation of technical terminology in which words have a special technical meaning not apparent in the words themselves.

CHAPTER SEVEN

CONTEXT AND CONCLUSIONS

I. THE POLITICAL CONTEXT OF EVALUATION

This monograph has attempted to clarify the concept of internal efficiency as it applies to the operation of educational institutions and systems and to characterize the nature of evaluation and the role of the evaluator in assessing educational efficiency issues in the development context. An explicit assumption of this presentation has been the belief that such evaluation efforts can only be appreciated within the political environment within which the evaluation takes place.

Shapiro (1985) has identified the following eight elements of the political context of evaluation:

- Decisions are usually a function of aggregate, rather than individual, behavior (Lindblom, 1965);
- Decisions are usually incremental in nature;
- Individuals calculate their positions in the political process in terms of individual goals and preferences;
- Evaluation information will be interpreted first in terms of its political, rather than its substantive, policy significance;
- Policy analysis data from evaluations will compete with the preferences of peers, political executives, and constituents for the decisionmaker's attention;
- Information gathering in the political process is a symbolic act designed to present the appearance of rationality, scientific objectivity, and authority in decisionmaking;
- Information is gathered more for the purpose of surveillance than for specific decisionmaking; and
- The information least likely to be utilized in decisionmaking is judgmental information (because of its susceptibility to misrepresentation).

This then is the quandary of the evaluative analysis of educational efficiency. To be done properly, it must make explicit the inherent subjectivity of goal selection and ordering of preferences and the evaluation process itself is one designed to culminate in

a judgment of value (of the program or project). Yet, if Shapiro's characterization of the political context is correct, as evaluation of educational efficiency approaches the standard of care and professionalism espoused here, the less likely it is that the evaluation will have an immediate effect in the political decisionmaking system.

The key term here, of course, is "immediate." Neither evaluation generally, nor the evaluation of educational efficiency specifically, can be relied upon to change projects or programs to which decisionmakers have a strong emotional or ideological commitment or in which they have a significant degree of personal and self-interest. Instead, evaluation and the concept of efficiency itself must depend upon the steady accumulation of evidence to demonstrate that they would have been of value in decisionmaking; only when this record is firmly established, can one expect decisionmakers to believe that evaluation and the efficiency concept will be of value as a critical factor. Shapiro's characterization is accepted as correct for the short run, but not necessarily as an appropriate description of the evolution of the use of data in the decisionmaking process.

The nature of the debate over the evolution of the role of evaluative data in decisionmaking can be indicated by two quotations. First, John Q. Wilson (1978) has noted that:

> When [organizations] use social science at all, it will be on an ad hoc, improvised, quick-and-dirty basis. A key official, needing to take a position, respond to a crisis, or support a view that is under challenge, will ask an assistant "to get me some facts."...social science is used as ammunition, not as a method, and the official's opponents will also use similar ammunition....there will be many shots fired, but few casualties except the truth.

C.E. Lindblom (1984) responds to Wilson by remarking that:

> Of course, social research has to be tied to positions, of course it is ammunition. But it is through the resulting challenge and counter challenge that usable truth often emerges; and, imperfect as the process is, indeed there is no feasible better alternative way of reaching an approximation to truth for social problem solving.

The Lindblom position is especially comforting for evaluators because the nature of the client/evaluator relationship breeds a natural bias toward advocacy rather than objectivity.

Unfortunately, in the developing world the market of ideas may be poorly developed, may operate under government constraints, or both. In evaluation activities in developing nations the evaluator is likely to be a monopolist rather than a

competitor. There may be no one else who knows enough about the project or program studied to offer a dissenting view and, unless access is granted to the evaluator's basic data, no debate on objective or interpretive grounds can be expected. In such a context, partisan, as opposed to more neutral and objective, evaluation may be especially harmful to the long-term credibility of the evaluative process itself. In an area of study as value-laden as that of educational efficiency analysis, the failure of an evaluator to make clear the assumptions and biases of the evaluative analysis can lead to a corruption of the policy debate and a diminution of the status of both research and evaluation activities.

II. SUMMARY AND CONCLUSIONS

The major conclusion to be derived from this survey is that both the efficiency concept and the evaluation process suffer from conceptual, definitional, and operational confusion. While in one sense both the concept and the process may be viewed as neutral tools of analysis, to become effective they must be operationalized in some specific value-laden manner. Efficiency in education means nothing without a political or some other subjective resolution of the goal matrix of education and the establishment of a value ordering of the goal combinations. Evaluation, in contrast to research, has as its product a value statement that embodies assumptions and interpretations that are inherently subjective and must go well beyond the objective findings of the valuation itself. The impact of this subjectivity is not to reduce the importance or utility of either the efficiency concept or the evaluation process. However, the subjectivity must be recognized and understood if planners and policymakers are to use efficiency evaluation results appropriately.

The lesson to be learned from the discussion presented here is that in the review of any evaluation of efficiency enhancement activities a planner or policymaker must be prepared to examine the assumptions that underlie the educational goal statements and to assess impartially the criteria, standards, and indicators used for both goals and the efficiency construct. To fail to do this is to delegate inappropriately a major degree of planning or policymaking authority to the evaluator. The products of an evaluation are meaningless unless understood in the context of the methodological and technical compromises that preceded them. One may be assured that such compromises are endemic to the efficiency evaluation process in schools or school systems.

The result of this paper's emphasis on the nature and limits of efficiency enhancement evaluation is not to provide justification for failing to conduct evaluations or for ignoring or for deprecating their results. Rather, the desired outcome is to heighten the level of discussion about such evaluation and to improve the probability that the final evaluative conclusions are not the product of a single individual's biases, implicit or explicit.

Another lesson to be derived from this discussion is that participation in the evaluation process should be inclusive and that the major stakeholders should play an active role in the debate over assumptions and specifications. It is necessary for the stakeholders in the evaluation to be identified clearly and to have their views incorporated--even if this requires multiple specifications of a single goal construct. Often this may result in an evaluation that suggests a particular project or program is valued positively for some stakeholders and negatively for others. This interpretive ambivalence is a reflection of the political reality of the educational process.

Evaluators have been reluctant to present their results in this manner because of the fear that such a report would be viewed as indicative of indecisiveness. The consumers of evaluations need to be educated to recognize that such plural evaluative outcomes are a sign of professionalism, not political equivocation.

While much of the preceding discussion would apply to any evaluation, this paper has attempted to show that the problems of subjectivity and multiplicity of stakeholders are compounded when dealing with the concept of educational efficiency enhancement. Although some researchers and evaluators have attempted to reduce the concept to a simple mechanistic metaphor based on the most readily obtainable data, efficiency enhancement in education remains an inherently fluid and diffuse process. It can be shaped to whatever form a particular individual or group desires.

Evaluators must recognize and restrain their own biases and attempt to identify and to articulate the values imposed on the evaluation by the stakeholders. Greater openness to the role of subjective judgments and values as determinants of the efficiency criteria and of the methodology of evaluation should, in the long run, advance the legitimacy of evaluation and increase the confidence of users in the products of the evaluation. Without such an approach, either planners and policymakers will remain at the mercy of the hidden agendas of evaluators, or more likely evaluators will find their results accepted only when the results agree with the preconceived assessments of the users. Neither situation justifies the time and expense required for a proper evaluation of educational efficiency issues.

The final point to be made here is one that was stressed earlier. The timing of an evaluation--when it occurs and how much time is allocated for it--is a much more critical issue than the commissioners of evaluative studies appear to realize. There needs to be an increased acceptance of evaluation as a continuous rather than an isolated or intermittent responsibility of management. This is especially crucial in the management of technical assistance activities where cross- cultural influences and other factors accentuate the normal constraints on management control. Assessment, formative evaluation, and summative evaluation should not exist in isolation. Rather, they should represent phases in a cumulative process whereby an understanding of a program's activities and of the values and preferences of program participants are increased. Without this condition, the evaluation of educational efficiency enhancement may prove to be as ineffective as so many of the educational efficiency enhancement efforts of the last two decades.

BIBLIOGRAPHY

BIBLIOGRAPHY

Babbie, E., The Practice of Social Research (fourth edition), Belmont, California: Wadsworth Publishing Co., 1986.

Balassa, B. "Policy Responses to External Shocks in Sub-Saharan African Countries." Journal of Policy Modeling, Vol. 5, No. 1, 1983, pp. 75- 105.

Balassa, B. "Adjustment Policies in Developing Countries: A Reassessment." World Development, Vol. 12, No. 9, 1984, pp. 955-972.

Baum, W.C. and S.M. Tolbert, Investing in Development: Lesson of World Bank Experience, Washington, D.C.: The World Bank, 1985.

Benrean, J. and N. Birdsall. "The Quality of Schooling: Quantity Alone is Misleading." American Economic Review, Vol. 73, No. 5, pp. 928- 946.

Benyahia, H. Education and Technological Innovations: Academic Performance and Economical Advantages. Montreal: Renoug, 1983.

Bridge, R.G., C.M. Judd, and P. Moock. The Determinants of Educational Outcomes: The Impact of Families, Peers, Teachers, and Schools. Cambridge, Ma.: Ballinger Publishing Co., 1979.

Brislin, R.W., W.J. Lonner, and R.M. Thorndike. Cross Cultural Research Methods, New York: John Wiley & Sons, 1973.

Brooke, N. and J. Oxenham. "The Influence of Certification and Selection on Teaching and Learning." In J. Oxenham (ed.), Education Versus Qualifications? London: George Allen and Unwin, 1984, pp. 147-175.

Burstein, L., "The Analysis of Multilevel Data in Educational Research and Evaluation," Chapter 4 in Review of Research in Education, Vol. 8, 198-.

Burstein, L., "The Role and Levels of Analysis in the Specification of Educational Effects," Chapter 3 in R. Dreeben and J.A. Thomas, The Analysis of Educational Productivity, Vol. I: Issues in Microanalysis, Cambridge, Mass.: Ballinger, 1980.

Chapman, D.W. and J. F. Carter, "Translation Procedures for the Cross- Cultural Use of Measurement Instruments," Educational Evaluation and Policy Analysis, 1, 3, 1979.

Chapman, D.W. and E.F. Kelly. "A Comparison of the Dimensions Used by Iranian and American Students in Rating Instruction," International Review of Education, 27, 1981, 41-60.

Chiswick, C.U. "The Impact of Educational Policy on Economic Development: Quantity, Quality, and Earnings of Labor." Economics of Education Review, Vol. 3, No. 2, 1984, pp. 121-130.

Cieutat, V.S. "Planning and Managing an Education Sector Assessment: Update," McLean, Va: Institute for International Research, 1986.

Cieutat, V.S. "Planning and Managing an Education Sector Assessment," Office of Science and Technology/Education, Washington, D.C.: United States Agency for International Development, 1983.

Clark, R.E. and A. Voogel, "Transfer of Training Principles for Instructional Design," Educational Communications and Technology Journal, 33, 2, 1985, 113-123.

Collier, P. and Lal, D. "Why Poor People Get Rich: Kenya, 1960-79." World Development, Vol. 12, No. 10, pp. 1007-1018.

Cronbach, L.S. and L. Furby, "How Should We Measure 'Change': Or Should We?," Psychological Bulletin, 74, 1, 1970, 68-80.

Cummings, W.K. The Conceptualization and Diffusion of an Experiment in Low-Cost Education: A Six-Nation Study. Ottawa: IRDC, 1984.

Dobson, D. and T.J. Cook, "Avoiding Type III Errors in Program Evaluation: Results from a Field Experiment," Evaluation and Program Planning, 3, 1980, 269-276.

Dominguez-Urosa, J. Efficiency and Other Indicators of Performance in Education Systems. Washington, D.C.: World Bank EDI Course Note Series, January, 1980.

Donohue, J.D. (ed.) Cost-Benefit Analysis and Project Design. Washington, D.C.: Training and Development Division, USAID, 1980.

Dore, R. "Human Capital Theory, The Diversity of Societies and the Problem of Quality in Education." Higher Education, Vol. 5, 1976, pp. 79-102.

Dore, R. and J. Oxenham. "Educational Reform and Selection for Employment--an Overview." In J. Oxenham, Education Versus Qualifications. London: George Allen and Unwin, 1984, pp. 3-40.

Eicher, J.C. Educational Costing and Financing in Developing Countries: Focus on Sub-Saharan Africa. Washington, D.C.: World Bank Staff Working Paper No. 655, 1985.

Eksterowicz, K. Contextual Effects and Relationships in Student Ratings of Secondary Instruction in New York State, Unpublished Doctoral Dissertation, State University of New York at Albany, Albany, New York, 1985.

Farhad-Motamed, "Internal-External Locus of Control of Iranian Fifth- grade Teachers and Students as Related to Student Social Class and Academic Achievement and Teacher Effectiveness," (Unpublished Master's Thesis) NIRT School of Cinema and Television, Tehran, Iran, 1977.

Foster, P. "The Educational Policies of Postcolonial States." In L. Anderson and D. Windham (eds.), Education and Development. Lexington, Mass.: Lexington Books, 1982, pp. 3-25.

Fullan M. and A. Pourfret, "Research on Curriculum and Instruction Implementation," Review of Educational Research, 47, 1, 1977, 335- 397.

Fuller, B. Raising School Quality in Developing Countries: What Investments Boost Learning? (Draft Version). Washington, D.C.: World Bank Education and Training Division, 1985.

Glass, G., "Standards and Criteria," Journal of Educational Measurement, 15, 4, 1978, 237-261.

Government of the Yemen Arab Republic, Education and Human Resources Sector Assessment in Yemen, Ministry of Education, Sanaa, Yemen, 1985.

Green, T.F. Predicting the Behavior of the Educational System. Syracuse, New York: Syracuse University Press, 1980.

Haddad, W. Educational and Economic Effects of Promotion and Repetition Practices. Washington, D.C.: World Bank Staff Working Paper No. 319, 1979.

Hall, G. and S.A. Loucks, "A Developmental Model for Determining Whether the Treatment Really is Implemented," Research and Development Center for Teacher Education, University of Texas at Austin, 1976.

Hanushek, E.A. A Reader's Guide to Educational Production Functions. New Haven, Conn.: Yale University Institute for Social and Policy Studies, Working Paper No. 798, December 1977.

Harris, C.W. "Some Problems in the Description of Change," Educational and Psychological Measurement, 22, 2, 1962, 303-319.

Hartley, M. and E. Swanson. "Achievement and Wastage: An Analysis of the Retention of Basic Skills in Primary Education." (Draft Version). World Bank Development Research Department, 1984.

Havelock, R. and A.M. Huberman. Solving Educational Problems: The Theory and Reality of Innovation in Developing Countries. New York: Praeger Press, 1978.

Heyneman, S.P. "Resource Availability, Equality, and Educational Opportunity Among Nations." In L. Anderson and D. Windham (eds.), Education and Development. Lexington, Ma.: Lexington Books, 1982, pp. 129-144.

Heyneman, S.P. "Improving the Quality of Education in Developing Countries." Education and Development: Views from the World Bank. Washington, D.C.: The World Bank, 1983

Heyneman, S.P., J.P. Farrell, and M.A. Sepulveda-Stuardo. Textbooks and Achievement: What We Know. Washington, D.C.: World Bank Staff Working Paper No. 298, 1978.

Heyneman, S.P. and D. Jamison. "Student Learning in Uganda: Textbook Availability and Other Factors." Comparative Education Review, Vol. 24, pp. 206-220.

Heyneman, S.P. and W.A. Loxley. "The Effect of Primary-School Quality on Academic Achievement Across Twenty-Nine High- and Low-Income Countries." American Journal of Sociology, Vol. 88, No. 6, 1983, pp. 1162-1194.

Hicks, N. and Anne Kubisch. "Cutting Government Expenditures in LDCs." Finance and Development, Vol. 21, No. 3, 1984, pp. 37-39.

Husen, T., L.J. Saha, and R. Noonan. Teacher Training and Student Achievement in Less Developed Countries. Washington, D.C.: World Bank Staff Working Paper No. 310, 1978.

International Monetary Fund. World Economic Outlook, April, 1985. Washington, D.C.: International Monetary Fund, 1985.

Jamison, D. Radio Education and Student Repetition in Nicaragua. Washington, D.C.: World Bank Reprint Series No. 91, 1978.

Johnson, Moritz, "Evaluation Reflections: The Locus of Value Judgments in Educational Program Evaluation," Studies in Educational Evaluation, 5, 1979, 109-122

Kelly, E.F., "Getting Value in Evaluation," School of Education, State University of New York at Albany, 1985.

Kelly, E.F., "The Role of Testing In American School Reform", in National Education Reform and New York State: A Report Card, Rockefeller Institute Conference Proceedings, Albany, N.Y.: State University of New York at Albany, 1985.

Kelly, E.F. "Evaluation of the Improved Efficiency of Learning Project in Liberia, Africa: Overview and Peculiar Problems," The Evaluation Consortium at Albany, School of Education, State University of New York at Albany, Albany, New York, 1984.

Kelly, E.F., "Horse Races, Time Trials, and Evaluation Designs: Implications for Future Evaluations of the Improved Efficiency of Learning Project," School of Education, Albany, N.Y.: State University of New York at Albany, 1984.

Kelly, E.F. "Evaluation: Issues and Practices," School of Education, Albany, New York: State University of New York at Albany, 1983.

Kelly, E.F. and D.W. Chapman. The Classroom Behavior Survey, Test Scoring and Evaluation Services, Syracuse University, Syracuse, New York, 1978.

Kemmerer, F. Towards Specification of a Student Time Supply Function Ph.D. Dissertation, University of Chicago, Chicago, Illinois, 1980.

Kemmerer, F. and J. Friend. "Strategies for and Costs of Disseminating the Radio Language Arts Project Throughout Kenya," Report of the Academy for Educational Development, Washington, D.C., March 1985.

Klees, S.J. and M.R. Suparman. An Evaluation of the Costs of PAMONG Schooling Alternatives in Indonesia. Washington, D.C.: Institute for International Research, 1984.

Levin, H.M. Cost-Effectiveness: A Primer. Beverly Hills: Sage Publications, 1983.

Levin, H.M. and L. Woo. "An Evaluation of the Costs of Computer-Assisted Instruction," Economics of Education Review, Vol. 1, No. 1, 1981, pp. 1-25.

Lewin, K. "Selection and Curriculum Reform," In J. Oxenham (ed.), Education Versus Qualifications? London: George Allen and Unwin, 1984, pp. 115-146.

Lindblom C.E. The Intelligence of Democracy, New York: The Free Press, 1965.

Lindblom C.E. "Who Needs What Social Research for Policy Making?" Rockefeller Institute Conference Proceedings State University of New York at Albany, No. 2, Fall, 1984.

Mincer, J. "Human Capital and Economic Growth," Economics of Education Review, Vol. 3, No. 3, 1984, pp. 195-205.

Mingat, A. and G. Psacharopoulos. "Financing Education in Sub-Saharan Africa," Finance and Development, Vol. 22, No. 1, 1985.

Monk, D.H. "Interdependences Among Educational Inputs and Resource Allocation in Classrooms," Economics of Education Review, Vol. 3, No. 1, 1984, pp. 65-73.

Morgan, R.M. "Mastery Learning and Programmed Instruction in Developing Countries," Educational Technology, October, 1973.

Nowicki, S. and B.R. Strickland. "A Locus of Control Scale for Children," Journal of Consulting and Clinical Psychology, 40, 1, 1973.

Organization for Economic and Development. Educational Planning: A Reappraisal, Paris, OECD, 1983.

Popham, J., Educational Evaluation, Englewood Cliffs, N.J.: Prentice Hall, 1975.

Psacharopoulos, G. and M. Woodhall. Education for Development: An Analysis of Investment Choices. New York: Oxford University press, 1985.

Psacharopoulos, G., K. Hichliffe, C. Dougherty, and R. Hollister. Manpower Issues in Educational Investment: A Consideration of Planning Processes and Technologies. Washington, D.C.: World Bank Staff Working Paper No. 624, 1983.

Robinson, B., "On Methodology for Education Sector Analysis," Washington, D.C.: United States Agency for International Development, 1973.

Sadler, R., "The Origins and Functions of Evaluative Criteria," Educational Theory, 35, 3, 1985, 285-297.

Shapiro, J. "Where We Are and Where We Need to Go," Educational Evaluation and Policy Analysis, Vol. 7, No. 3, Fall 1985.

Simmons, J. How Effective is Schooling in Promoting Learning? A Review of the Research. Washington, D.C.: World Bank Staff Working Paper No. 200, 1975.

Stake, R.E. "Testing Hazards in Performance Contracting," Phi Delta Kappa, June, 1971, 583-589.

Stake, R.E. "The Countenance of Educational Evaluation," Teacher's College Record, 68, 1967. 523-540.

Stanley, J. "General and Special Formulas for Reliability of Differences," Journal of Educational Measurement, 4, 1967, 249-252.

Stufflebeam, D.L. et. al., Educational Evaluation and Decision Making, Itasca, Ill: Peacock, 1971.

Tan, J.P., K.H. Lee, and A. Mingat. User Charges for Education: The Ability and Willingness to Pay in Malawi, Washington, D.C.: World Bank Staff Working Paper No. 661, 1984.

Taylor, D. and R. Obudho. The Computer and Africa, New York: Praeger Publishers, Inc., 1977.

Thiagarajan, S.P. Appropriate Educational Technology for Developing Nations: Low Cost Learning Systems, Bloomington, Indiana: Institute for International Research, 1984.

Thomas, J.A. Resource Allocation in Classrooms, Chicago: University of Chicago Educational Finance and Productivity Center, 1977.

Tuckman, H.U., T.F. Nas, and J.S. Cladwell. "The Effectiveness of Instructional Radio in a Developing Country Context," Middle East Technical University Studies in Development, Vol. 10, No 1., 1983, pp. 45-64.

Walberg, H.J. "Improving the Productivity of America's Schools," Educational Leadership, Vol. 41, No. 8, 1984, pp. 19-26.

Watson, K. (ed.) Education in the Third World. London: Croom Helm, 1982.

Weiss, C.H. Evaluation Research, Prentice Hall: Englewood Cliffs, N.J., 1972.

Wells, S. Instructional Technology in Developing Countries: Decision Making Processes in Education. New York: Praeger Publishers, Inc., 1976.

Wiles, P. "The Correlation between Education and Earnings: The External-Test-Not-Content Hypothesis," Higher Education

Wilson, J.Q. "Social Science and Public Policy," in L.E. Lynn, Jr. (ed.), Knowledge and Policy (Washington, D.C.: National Academy of Science, 1978).

Windham, D.M. "Micro-Educational Decisions as a Basis for Macro- Educational Planning," In H.N. Weiler (ed.), Educational Planning and Social Change. Paris: International Institute for Educational Planning, 1980, pp. 107-121.

Windham, D.M. "The Dilemma of Educational Planning," In L. Anderson and D. Windham (eds.), Education and Development. Lexington, Ma.: Lexington Books, 1982, pp. 159.174.

Windham, D.M. "Cost Issues in Liberian Improved Efficiency of Learning Project." Report prepared for the Liberian Ministry of Education and the Institute for International Research, January, 1983.

Windham, D.M., "The Relative Cost-Effectiveness of the Liberian Improved Efficiency of Learning Project," Report prepared for the Liberian Ministry of Education and the Institute for International Research, June, 1983.

Windham, D.M., "Cost Issues in the Dissemination of the Liberian 'Improved Efficiency of Learning' Project," Report prepared for the Liberian Ministry of Education and the Institute for International Research, June, 1983.

Wolfe, B.L. and J.R. Behrman. "Who is Schooled in Developing Countries? The Roles of Income, Parental Schooling, Sex, Residence, and Family Size," Economics of Education Review 3, 3, 1984, pp. 231-245.

Wolff, L., Controlling the Costs of Education in Eastern Africa: A Review of Data, Issues, and Policies. Washington, D.C.: World Bank Working Paper No. 702, 1985.

The World Bank, Education Sector Policy Paper. Washington, D.C.: World Bank, 1980.

The World Bank, World Development Report, 1985. Washington, D.C.: The World Bank, 1985