

**BIBLIOGRAPHIC DATA SHEET**

1. CONTROL NUMBER

PN-AAJ-329

2. SUBJECT CLASSIFICATION (100)

JC70-0000-0000

## 3. TITLE AND SUBTITLE (100)

Educational outcomes in less developed countries: analysis and evaluation tools; Chapters IX-XX

## 4. PERSONAL AUTHORS (100)

## 5. CORPORATE AUTHORS (101)

Georgetown Univ. Public Services Laboratory; Haile Sellassie I Univ.

## 6. DOCUMENT DATE (110)

1973

## 7. NUMBER OF PAGES (120)

193p.

## 8. ARC NUMBER (170)

370.18.G351

## 9. REFERENCE ORGANIZATION (130)

Georgetown

## 10. SUPPLEMENTARY NOTES (500)

(Cosponsored by the Institute of Development Research, Haile Sellassie I Univ.)

(Chapters I-VIII, 198p. :PN-AAJ-328 )

## 11. ABSTRACT (950)

## 12. DESCRIPTORS (920)

Education

Measurement

Achievement tests

Education for development

Teacher training

Attitude surveys

Educational planning

Tests

Evaluation

Curriculum development

## 13. PROJECT NUMBER (150)

931099400

## 14. CONTRACT NO.(140)

AID/CM/ta-C-73-8

## 15. CONTRACT TYPE (140)

## 16. TYPE OF DOCUMENT (160)

370.18  
G351  
v.2

PN-AAJ-329

**educational outcomes  
in less developed  
countries:  
analysis and  
evaluation tools**

**HAILE SELASSIE I UNIVERSITY  
GEORGETOWN UNIVERSITY**

## CHAPTER IX

### MEASURING ACADEMIC ACHIEVEMENT

Thus far we have considered educational outputs in the framework of the developmental objectives of nations--economic growth, occupational manpower needs, and human development. Planners have usually turned to general statistical indicators of educational needs in making decisions about resource allocation. Recently, in developed countries, attention has been directed toward the use of more specific indicators--assessments of student attainment--in defining system outputs and future needs. Assessment of student achievement can specify in detail the amount of knowledge acquired by groups of students in distinct geographic regions. It will inform planners, for example, that a certain percentage of students in a given grade can read a selected text meeting pre-stated functional criteria of speed and comprehension. In contrast, a typical indicator currently available for planning might be the proportion of the population above a certain age which is literate. In this chapter and the next, we look at existing and potential measures of student educational outcomes in the developing nations.

The first part of this chapter defines two general types of assessment and reviews the current state of educational assessment in developing countries--the characteristics of national examinations, the

examination reform movement, and the growth of standardized testing in developing countries. The second part focuses on decisions and problems confronting Education Ministers instituting standardized testing programs. Finally, we look at some of the broader issues implied by large scale testing programs.

### I. Maximum and Typical Performance

Educators have traditionally thought of learning outcomes as occurring in three broad categories:

Cognitive learning is the process by which we come to understand and organize the world as we perceive it. Cognitive skills are abilities such as acquiring knowledge, understanding phenomena, solving problems, creating new forms and ideas, and communicating effectively.

Affective learning involves the growth and maturation of aspects of the personality such as character and temperament. Many of the characteristics labeled "affective" in education are partly cognitive in nature. A purely affective characteristic is a mood or temporary emotional state. More lasting characteristics such as values or attitudes are largely cognitive, but have important affective consequences.

Psycho-motor, or Perceptual-Motor, learning is a crucial dimension of child development before age seven. This domain encompasses skills such as balance, left-right discrimination, hand-eye coordination, agility, endurance, flexibility, strength, relaxation, perception of body image.

However, the more that we learn about learning and development the more these categories seem to overlap; abilities in each category are strongly interrelated. Certainly skills in each domain affect the

development of faculties in the others. Much recent investigation has emphasized the mediating mechanisms in learning which have both cognitive and noncognitive components. These include motivation to achieve, expectancy of success, self esteem, cognitive style, and inhibitors such as anxiety. These characteristics are both inputs and outcomes of education and have a profound effect on achievement.

A useful classification of outcomes from an assessment perspective distinguishes between capacity measures and those bearing on expected performance. One type of assessment is measurement of what a student can do--the maximum behavior or ability. This type of measurement may be concerned with what a student can learn (aptitude testing) or with the expression of that capacity in what a student has learned (achievement testing). This chapter treats the measurement of these indicators of maximum ability, or academic outcomes.

The second general type of assessment is the determination of characteristics which affect what a student will do--the typical behavior that can be expected. Appraisal of personality characteristics and the mediating mechanisms fall into this category, and will be discussed in Chapter X. Current educational measurement research approaches summative and predictive assessment as a combination of aptitude/achievement testing and some optimal group of the second type of measures.

### Current Examination Practices and their Modernization

Educational policy makers in developing nations are now examining existing assessment practice in the light of national goals, as well as the current state of the art of testing. A corresponding inquiry into the connections between curriculum, assessment and educational objectives--

relationships long neglected--is gathering momentum. Worldwide, there has been a surge of interest in the reform of examination practices inherited from European traditions. The national external examination on the English or French model is now seen in many developing countries as largely irrelevant to present national needs and realities.

The external examination is a written essay or oral test based largely on the recitation of memorized material. It evolved in post-Renaissance Europe as a mechanism of control over the schools by subsidizing agencies (church or state) and was exported to the developing world with the European educational structure. Today, the external examination is still designed and controlled by agencies outside the school and often beyond the influence of teachers or administrators. The examinations have typically tended to function quite independently of the educational policies guiding the schools. Yet, because they determine admission to higher education, the examinations exert an enormous sway over teaching. Studies have discounted the validity of European external examinations as selection devices, noting scoring variations between examiners, variations by the same examiner, fluctuations in level of difficulty from year to year and inadequate sampling of subject matter. The examinations are even less effectual in developing nations. The school leaving examination in a developing country usually controls the content of instruction, dominates the attention of students and teachers, and causes widespread anxiety and frustration.

The need for basic changes is clear. The traditional examination does not produce comparable and replicable data about what students know and what they do not know. It does not assess

achievement, but measures linguistic fluency and certain elusive qualities of the intellect associated with a humanistic and classical education. The continued emphasis on qualities associated with European classical traditions in the developing countries has probably contributed to the exaggerated verbalism and lack of practical knowledge of some college graduates noted by a few Latin American and South Asian observers.

Even when national examinations are revised, as they have been in the last decade in many countries, usually they are not subjected to the rigorous analysis required to insure validity. A case in point is the Ethiopian School Leaving Certificate Examination (ESLCE) which has been given to students in the last year of academic secondary school for the last 19 years. The ESLCE is prepared, administered and scored by subject matter specialists associated with the Haile Sellassie I University. A candidate who takes the ESLCE is successful if he attains five passes at C level, including the Amharic Language Examination, the English Language Examination, and the Mathematics Examination. There is no limit to the number of times a candidate can take the ESLC Examination, and the time allocated for each examination varies.

The ESLCE has long been controversial. The percentage of those passing the exam sank from 89 percent in 1950 to 20 percent in 1966, causing widespread public dissatisfaction and disputes between teachers and examination officials over the cause of the high failure rate. In 1967, a number of commissions were set up to examine the content of the examinations and the high school curriculum. Participants declared that contributing to the failure rate were promotion of students without adequate skills from the lower grades, poor instruction by high school teachers, and inadequate student preparation for the exam.

Teachers claimed the exams had ambiguous questions, limited coverage of subject matter, and allowed too little time for answering questions.

At the conclusion of a five day session, the Commissions recommended enrichment of the curriculum and revisions of the ESLCE, including the substitution of objective test items for some essay questions. But significant reduction has been made in the number of failures, which varied between 83 percent and 74 percent between 1969 and 1973.

The issue to be addressed is whether the revised ESLCE is valid (does it accurately test the knowledge, skills, and understanding that a high school graduate should have) and reliable (are the results consistent)? Preliminary findings have indicated that the predictive validity of the ESLCE for university performance has been low and the reliability of the test has never been established. Yet, the examination is used as the yardstick for determining "successful" completion of secondary school.

The University Testing Center at HSIU has recommended further work in the evaluation of validity and reliability of the tests, and that the teacher's appraisal also be taken into account when assessing candidates for admission.

#### Standardized Testing: One Direction of Reform

The widespread examination reform movement, stimulated by extensive unhappiness with existing examination practices, will have a great effect on future collection of data on educational outcomes. Proposed solutions range from the abolition of all exams to the substitution of standardized exams. Some educators are seeking to develop new essay-type exams with shorter questions, which would provide a better sampling of subject

matter. The major trend, however, is toward objective standardized tests, which have been used in developing countries in the past for admission to foreign universities and for certain types of vocational selection. A standardized test is one which is repeatable for various populations at different times because the test content (though not the exact items), procedure, and scoring are fixed.

A few examples will illustrate uses of standardized tests in developing nations. Singapore has instituted standardized testing programs at the end of the 6th year of primary school, the 4th year of secondary school, and the 2nd year of postsecondary school. These are designed by the Examinations Division of the Ministry of Education, working with the Cambridge Overseas Examinations Syndicate. The Primary School Leaving Examination covers two languages (choices include English, Chinese, Malay or Tamil), Mathematics and Science.

In Vietnam, the Ministry of Education, Culture and Youth abolished the 11th grade baccalaureate in 1973. The Ministry is now designing standardized tests, to be administered to high school students completing the 12th grade in 1974. These tests will replace the 12th grade baccalaureat.

In the Philippines, a standardized college entrance exam (CET) was first administered on a voluntary basis two years ago. It was produced by a private Phillipine organization which has now contracted with the government to develop a national college entrance exam (NCEE) which will be required for admission to all postsecondary schools. The NCEE will have two parts, a subject proficiency section measuring English, mathematics and science; and a mental ability section consisting of tests

in verbal analysis, number and letter series, word-number relations, and abstract reasoning.

Indonesia provides an interesting example of the evolution of testing reform encompassing both localization of control and national data gathering. In the 1950's, education officials concerned with providing all children an education at the elementary school level concluded that the current examinations and curriculum were unsuitable for a democratic educational system. The uniform curricula did not reflect the cultural and geographic diversity of Indonesia, and the harsh external exams were not functional for selection. A state commission set up in 1958 to study these problems recommended that external examinations be abolished following a nationwide upgrading of teacher skills in testing and evaluation. A central board of examinations was established to advise and direct local examination boards in the management of the exams, and a schedule was set for the step-by-step decentralization of the examination organization. After 1969 all exams were produced either by teachers or, in some parts of the country, by provincial committees. Beginning in 1973, each school developed its own examinations.

At the same time, the Ministry developed a standardized testing program for educational planning purposes. These tests will be given to all children in grades one through 12, and cover mathematics, science, social science, and Indonesian languages. The questions are short answer and include the cognitive areas of knowledge, comprehension, and application. Two functions of testing are thus clearly separated: the evaluative and guidance function for students and teachers, and the assessment of education outcomes for educational planners.

## II. Measuring Educational Outcomes with Standardized Tests

The necessity for assessment of educational needs and outcomes is common to all countries. LDC's, however, have a smaller margin for experimentation in the allocation of resources. It is in these countries, where the need for a close analysis of investments in education is greatest, that the data base for decision making is least complete and accurate. LDC's spend a high proportion of their national budgets on education and clearly cannot afford to continue increasing the overall investment in education at the expense of other social programs.

Educational expenditures should be differentiated in function and payoff consequences. Planners must be able to weigh the relative merits of program investments with regard to quantitative increase in education, needed changes in teacher recruitment and training, needed changes in curriculum, and the diversification of the institutional role of education in national life. Viable national planning in these areas can be assisted by measurements of:

- (1) the numbers of people to reach each level of education annually;
- (2) the availability of the resources for educational development: teachers, facilities, curriculum and teaching methodology and technology, financial resources; and
- (3) the abilities of individuals, inside and outside the formal educational system, in cognitive, vocational, motivational and other spheres.

It is the third type of measurement that is most deficient in developing countries (and in developed countries); it is this type of data which standardized testing can supply. Specific data on the acquired skills and educational deficits of demographically differentiated groups of students can be collected on a continuing basis for systems analysis and planning. Furthermore, basic skills of large numbers of people outside the formal system--adults and youth who never entered school or dropped out--can be assessed for vocational placement or to allow them to reenter formal schooling at the proper level. Finally, standardized testing is appropriate for certain types of nonformal programs. If technical skills are taught in diverse ways, certificates of program completion may have little meaning in the market place. Although practical tests will continue to be important, employers have learned that a well designed standardized test, validated to criterion skills determined by the job specifications, can assess large numbers of applicants economically (see Chapter VI on work skill assessment).

#### Instituting a Standardized Testing Program

Implementation of standardized testing is a sizable task. The Education Minister or official must first decide whether to (A) use an existing test--published in another country or by the International Association for the Evaluation of Education--or (B) design a new test, for that country alone or cooperatively with other countries. Test construction is a considerable undertaking, as it is costly and requires specialized expertise and time for development. However, the investment can be justified for developing countries on the basis of long-term planning and educational benefits. It is probable that the information

requirements for policy making can be fulfilled only through developing testing instruments in the context of the particular social needs and problems of a country or region.

The testing options listed above will be examined in light of necessary procedures and policy considerations.

#### A. Use of Existing Tests

Tests developed in other countries are rarely appropriate for use in developing countries without some adaptation. Even where the language of the test (usually English or French) is a desired feature, some adjustment for reading level may be needed. An exception might be the International Assessment instruments, which were designed for cross national use and have been administered in some developing nations. The advantage of the IEA tests is that they were painstakingly constructed and extensively validated. However, it must be remembered that the International Assessment effort was initiated to analyze the effects of various in-school and out-of-school variables on learning. The research orientation of these tests may not coincide with the outcome information needs of a country.

The advantages of adaptation of existing tests are:

- (1) the relative speed and economy of the process of adaptation in comparison with test construction; and
- (2) a standardized test has little value if it has not gone through the processes of standardization and validation discussed later in this chapter. These procedures may present great problems to developing countries in terms of knowledgeable personnel, data processing systems, and financial support, and hence use of existing tests, previously validated, may offer a solution to these problems.

However, the problems with adaptation may outweigh the initial convenience and economy. There are several basic considerations in test adaptation: language, content, mechanics, and rationale.

Translation is a self-explanatory task, but may pose difficulties if the test contains vernacular phrases or proverbs. Content adjustments are routinely made by test adaptors on a superficial level (for example, changing monetary or measurement systems), but other issues of content are often ignored. The accuracy of a test is greater if the applicants are familiar with the examples and illustrations used. A common error has been to assume, in a vocational aptitude test, that an item measuring a certain type of mechanical aptitude must be presented in terms of the western machine or tool, rather than a more familiar tool from the local environment. The test is intended to measure the target skill; knowledge of the machine to be used on the job belongs to the realm of job training.

Testing mechanics must be modified by the test administrator according to circumstance. Test-taking is a skill improved by practice and familiarity. A separate answer sheet sometimes causes confusion for groups who are not accustomed to standardized tests. Another element is timing: cross-cultural research shows that people have very different senses of time. The time limit prescribed by a standardized test may have to be extended or eliminated to be fair to some groups of subjects. The instruction to "work as quickly and as accurately as you can" involves a complex type of judgment which is particularly alien to many cultures.

The test rationale deserves careful consideration. Every test is designed within a particular cultural context, and is usually validated against practical criteria in that particular culture. The consequences of translating the test to another culture can be evaluated only in terms

of the purpose of the test. Poor facility in the language, slow work habits and lack of abstract thinking will affect test scores, but may also affect the criterion behavior which the test was designed to measure (for example, in a vocational aptitude test). In other cases this is not necessarily true.

There are at least six levels of adjustment necessary for adaptation of instructional materials and examinations:

- (1) translation of language;
- (2) changing the vocabulary for the proper reading level;
- (3) changing illustrations and photographs to reflect local conditions;
- (4) adjusting procedures implied or specified to match the expectations and experience of the learners;
- (5) adjusting the content to reflect the local culture and life style; and
- (6) accommodating the learning styles of the students.

The last three are complex and require intimate knowledge of the culture. These considerations underline the necessity for nations to acquire trained psychometricians with the technical and cultural requisites for analyzing and designing tests for the functional requirements of the educational system.

The Haile Sellassie I University Testing Center in Ethiopia has had extensive experience in using and adapting tests such as the Differential Aptitude Test, the Davis Reading Test, and the Amharic Otis 100 (adapted from the Otis Quick Scoring Mental Ability Test).

The following are some of the conclusions drawn from the experiences of the Testing Center with respect to the assessment of educational

outcomes with cognitive measures adapted from other countries:

- (1) the studies conducted demonstrated that cultural factors greatly affect the ability of test items to elicit the intended responses; comprehensible test items must be constructed in the light of the examinee's cultural background;
- (2) since lack of familiarity with the procedures and types of tests poses a threat to the validity of the instrument and reliability of the test results, it is essential that the necessary information about tests be given to examinees ahead of time;
- (3) the indiscriminate use of standardized tests from developed nations is likely to be loaded with sampling error in connection with student abilities due to differences in the populations; and
- (4) validation (the process of insuring accuracy) for the criterion behavior toward which the test is directed must be undertaken in the adapting country. Before tests are used as predictors of success for the desired criterion, a scientifically designed pilot must proceed to verify that the tests actually predict the intended performance of the individual. This aspect of test development needs greater emphasis for better validity and reliability of the test results.

B. Development of a New Test

Designing new standardized tests to measure educational outcomes has several distinct advantages.

- (1) The effect of cultural influences on test performance may be pronounced. An adapted test may not be suited to the population because of built-in assumptions about cognitive skills, familiarity with concepts or phenomena, or cognitive style (see Chapter X). Necessary revisions in test mechanics and administration may be extensive.
- (2) A new test can be designed specifically for the curriculum and educational objectives of the formal school system. Tailor-made tests are essential for assessment of objectives in non-formal programs.
- (3) The test can be designed to supply the particular kinds of outcome data for the particular populations (grade levels) required by planners.
- (4) The process of test development is an opportunity to increase levels of awareness and expertise with educational measurement among teachers, administrators, and educational researchers.

The high cost of test development can be minimized by a regional pooling of financial and manpower resources by several geographically and culturally related countries. Measurement experts, teachers, educational psychologists, anthropologists and educational planners working together in a centralized research laboratory can generate common examinations and specialized tests for the participating countries.

The West African Examinations Council is an example of such an arrangement.

If a central problem is the lack of measurement specialists in a country or region, a solution is to bring testing experts from abroad to coordinate the test development project. The contract with the consulting organization can specify training in measurement and test design for personnel in the Ministry of Education and workshops for local administrators and teachers in measurement, test item writing and evaluation. It has proven important, in the developed countries, to have teachers involved in some phase of test design to promote a supportive attitude toward the procedure and rationale of standardized testing.

In this section we look at some of the basic decision areas facing an Education Minister or regional policy group engaged in implementing a testing program. These include: (1) defining the purpose of the test; (2) determining grade level and timing of test administration; (3) emphasizing aptitude, achievement testing or a combination of the two; (4) choosing norm referenced or criterion-referenced instruments; (5) defining the basic educational objectives and content to be covered; and (6) test construction and validation.

#### 1. Defining the Purpose of the Test

The reason for giving a standardized test must be clearly understood by test administrators, school personnel, and the students. In countries where standardized testing of student achievement is common, dissatisfaction sometimes results from imprecise definition of the goals of the test, which in turn determine the appropriate methodology, the analysis

of the data, and the fairness of the practical consequences. A test which is designed for one purpose is usually inappropriate and often unfair when used for a different purpose.

Educational measurement provides either needs assessment or treatment adequacy assessment data. Needs assessment is identifying the specific goals toward which education should be directed. Sometimes needs assessment is used to identify specific learner needs in terms of understood goals, for prescriptive or placement purposes. Treatment adequacy assessment refers to evaluation of the educational process; it is either formative, which is on-going testing during the instructional program, or summative, which takes place at the end of the program to judge its success. These three types of testing yield antecedent, formative, and outcome data. In some cases, tests at an intermediate level may be used both as a summative evaluation of a completed program and for needs assessment for treatment planning for the next level. In principle, however, the purposes are distinct, and ideally the tests to serve each should be tailored to the purpose.

Some specific uses for outcome measurement for different groups are noted below. The uses are discussed more fully in Chapter XVII.

- (1) educational research: contributions to the development of educational learning theory;
- (2) students and teachers: diagnosis of the learner's strengths and weaknesses for individual guidance and evaluation of student progress;
- (3) administrators: evaluation of educational programs, selection of students for the next level of education,

- distribution of students in different types of programs;
- (4) employers: certification of competence in required skills and knowledge; and
  - (5) educational planners: evaluation of the quality, efficiency and adequacy of the educational system as a whole.

Although one test can serve more than one function (i.e., provide data on individual student achievement for students and teachers as well as gross data on achievement for national planners), it must be carefully designed in order to do so. Tests cannot accomplish many goals indiscriminately; they must be designed for specific purposes, in the context of the particular situation and goals of a country. This constitutes one of the difficulties of transplanting tests designed for other countries.

## 2. Grade Level and Timing of Tests

An early decision which must be made is which grades are to be tested and with what frequency. A logical schedule is testing at the end of each stage: lower and upper primary, intermediate, and secondary. Due to the high wastage rate in many countries, testing should begin fairly early in order to get accurate outcome data. Another more costly plan is to test every other grade, beginning in the second grade. This may be appropriate where elementary grades are grouped in units of two grades each, as in the Arab States. On the other hand, testing too often--for example, at the end of each school year for admission to the next grade--can have damaging effects. Again, the certification and system evaluation functions of testing must be kept clearly differentiated.

Yearly testing for admission to the next grade is a practice which should be modified. Yearly tests for promotion are expensive, may cause extensive anxiety, and have a low predictive value for young children. Learning theorists have discovered that children develop cognitively at different rates. A third grader with good reading and personal-social skills may have difficulty with math and science--is it productive for the child to repeat third grade? The decision should be made by the teacher and guidance counselor.

The lock-step approach to instruction and individual assessment handicaps children who develop more slowly--a child should not be labeled a failure after one or two years of school. In the early years of schooling, it is important to balance standardized tests with more subjective teacher evaluations. At this stage, standardized tests are more useful for group assessment than individual assessment.

### 3. Aptitude and Achievement Testing

Educators developing a new system of assessment to replace the traditional examination have a choice between (A) aptitude testing; (B) a combination of aptitude and achievement testing; and (C) achievement testing alone.

Our primary concern in this report is the assessment of incremental gains by students--what might be called the value-added concept in measurement. This clearly requires achievement testing on a regular basis so that longitudinal data can be collected.

However, planners in some countries may wish to implement standardized testing not merely for planning purposes, as Indonesia has done, but also for purposes of selection for high levels of education. Combining

two functions of assessment--for planning purposes and for selection--cannot be taken lightly, since the former implies achievement testing and the latter has in the past implied aptitude testing.

A brief review of the advantages and disadvantages of the options listed above in the light of both these functions follows.

i. Aptitude Testing

Aptitude tests attempt to measure innate ability and have been favored as a means for predicting future performance. A few key areas, such as verbal and mathematical ability, can be emphasized and therefore the aptitude test may be more economical to develop than an achievement test. Another reason aptitude tests are used for selection is that the aptitude test theoretically circumvents problems of unequal quality and differing curricula in schools and consequent unequal academic achievement of students.

The disadvantages of aptitude testing are two-fold. First, it does not yield systematic output data for system evaluation and planning purposes on the national level. Secondly, it is extremely difficult to design a scholastic aptitude test that does not to some degree measure achievement.

Both general and specific aptitude tests are subject to the second criticism. General aptitude or I.Q. tests purporting to measure intelligence, of basic ability to learn, have been the subject of much controversy for decades. Factors such as socioeconomic status (and hence greater opportunity to learn) are so highly correlated with intelligence test scores that the tests have acquired some disrepute, especially in a cross-cultural context. The obvious deficiencies of

some intelligence tests spurred an effort to create "culture fair" tests, which would substitute abstract concepts or universal pictorial symbols for culturally-linked content. Further research revealed that even geometric shapes are perceived differently in different cultures. Also, the content left after eliminating all culturally-differentiating material may be too trivial to be valid for any practical criterion.

Specific scholastic aptitude tests are based on the theory that mastery of a specific skill results not from generalized ability but from particular abilities which can be isolated. These mental abilities, or traits, were identified by factor analysis of responses to test items by large numbers of people. Scholastic aptitude tests are discrete tests generated from the clusters of related items resulting from the factor analysis. (There are also non-scholastic aptitude tests which are widely used: skill tests for abilities such as hand-eye coordination, and vocational aptitude tests simulating particular job skills).

Some researchers objected to this model, in which the mind consists of independent constituents which can be separately measured, proposing an alternate model of the mind as an organic structure of complex inter-relationships. They maintain that specific cognitive abilities, like performing mathematical operations, consist mainly of the command of the knowledge pertaining to the process; and that general ability is primarily clusters of closely related specific abilities. Therefore, much of what has been seen as innate ability would actually be attributable to previous learning experiences.

The validity and usefulness of the mental trait theory to explain individual differences in learning is yet to be resolved. Aptitude

tests have a diagnostic function--they can be used to prescribe appropriate learning environments for individual students (for example, programmed instruction, lecture method or unstructured environment). However, it seems clear that most aptitude tests to some extent measure achievement, and since they have not been designed to do so, they cannot be used for achievement assessment.

ii. Combining Aptitude and Achievement Testing

It has been proposed that the two types of measurement be implemented in combination. Schwarz (1972) has argued for secondary school selection in developing nations by means of a two-tiered process. First, an achievement test would screen out students who had not mastered skills necessary for success in the secondary course; then a scholastic aptitude test would be used as the predictive instrument to identify the best candidates for secondary school. His rationale is based on the assumption of achievement testing that the past schooling experience of test takers is approximately the same; since this is not true in developing countries, the achievement test alone is inequitable.

This proposal does guarantee that students selected for secondary school would possess the requisite skills for progress, but the use of an aptitude test as the ultimate predictor has the same drawbacks listed in the previous section: lack of outcome data for planning purposes and the difficulty of designing a "pure" scholastic aptitude test not actually measuring achievement. Also, the ongoing controversy over whether intelligence is attributable more to hereditary or environmental factors suggests that there is no way to know to what degree an aptitude test may unfairly handicap part of a population.

The only way to assure detailed outcome data on educational outcomes while using aptitude testing for selection is to implement both testing programs simultaneously, which is not an economical procedure.

### iii. Achievement Testing

Achievement tests are straight-forward measures of educational outcomes. For students and teachers, achievement tests provide feedback information on progress and problem areas; for administrators, they allow comparative evaluations of programs or schools within a region. For the national planner, whose problem is to conduct educational assessment which will evaluate the effectiveness of the entire system, achievement testing is the logical tool. Valid and relevant assessment entails objectives-based exams with adequate coverage. Coverage must concern itself not merely with sufficient sampling of curricula, but with expected and desired behavior in terms of national goals; otherwise, objective assessment is open to the criticism of trivializing the aims of education. Aptitude tests, which are more selective in content and are based on the premise of prediction rather than of assessment, clearly will not satisfy the needs of planners or administrators for system evaluation. Achievement tests appear to have the capacity to perform both functions.

The difficulties imposed by differing curricula and quality of school instruction may be partially overcome by some recent technical developments in test construction, particularly the criterion-referenced test discussed below.

#### 4. Selection of Instruments

Most standardized tests in use today are norm referenced, which means that they are designed to show how an individual scores in relation to other individuals taking the same test. The norm-referenced test has been used in competitive situations where a few individuals must be chosen from many for a limited number of educational or vocational positions, as is usually the case in developing nations.

In recent years testing experts have developed the criterion-reference test (or domain-reference test), which is designed to measure a student's ability to perform (or mastery of) a particular skill or concept. The significance of the criterion-referenced test is that it gives a great deal more information about what a student (or a group of students) actually knows or can do, but less information for use in comparing individual students to one another. However, if desired, the criterion-referenced test can be designed to give normative information. Norm-referenced tests are designed to produce variability; a proportion of the students will always fail the test. Statistical considerations, as well as content validity, influence the inclusion of test items. Criterion-referenced tests, which must be constructed around specific learner objectives, can theoretically reveal that 100 percent of the students mastered the instructional objectives.

For nearly every function of testing, criterion-referenced tests provide better information. Where the emphasis is on the assessment of quality of education, they are clearly superior. It is also true that they are typically more expensive to produce than norm-referenced tests. Work is now proceeding on the technical problems of criterion-referencing.

As models for these tests become more readily available, they will become more economical and more applicable for selection. A test designed to assess mastery of skills needed to succeed in secondary school would be a better predictor of academic success in secondary school than a norm-referenced test based on the elementary school curriculum. The norm-referenced test is more likely to yield misleading estimates of performance and to reflect existing social status in selection. Following a criterion referenced screening test, candidates can be selected from a pool of qualified candidates according to other criteria (geographic, for example) resulting in greater access to education for non-privileged students. Other factors which can be utilized as predictors of academic success, such as achievement motivation, are discussed in Chapter X.

In most cases, a carefully designed and administered standardized achievement test, possibly combined with other methods of evaluation, will probably be the best way to select students for further schooling.

##### 5. Defining Objectives and Test Content

One reason for caution in adapting existing instruments has been discussed: a good test, like an effective curriculum, reflects the values and educational objectives of the society in which it is given. Ultimately the aims of a society are the basis for evaluation of the society's educational system.

Objectives-centered assessment entails close analysis of the cognitive skills implied by educational goals. There are three very general types of goals: that students acquire the necessary knowledge and skills to (a) live happily and successfully; (b) fill social roles, including

the nation's manpower goals; and (c) preserve and develop the traditions and values of the society and humanity at large. The elaboration of these goals into specific skills and knowledge is the foundation of curriculum and evaluation design.

More specifically, Benjamin Bloom (1956) has classified educational objectives in the cognitive domain in a hierarchical taxonomy with six ascending levels

- (1) knowledge of facts, concepts, generalizations and methods;
- (2) comprehension, interpretation, and extrapolation of information;
- (3) application of principles;
- (4) analysis of material, recognizing relations and organization;
- (5) synthesis of parts to produce a new communication; and
- (6) evaluation or judgment.

A taxonomy of this type can be utilized by the designers of standardized tests to insure that their tests incorporate the higher orders of cognitive learning and not merely the lowest level of repetition of facts. Interpretation and reporting of test scores determines the degree to which program evaluation is possible. Tests should report a separate score for each objective rather than an average score.

Evaluation has a feedback effect on the instructional process. If examinations determine acceptance to higher levels of education, they tend to control the content of curriculum. It is counterproductive to promote national priorities in curriculum while neglecting the relationship between the content of examinations and national goals.

## 6. Test Construction

The construction of a standardized test involves many procedures, including:

- (1) writing of test specifications: a complete and explicit listing of all the characteristics of the test, which means resolving all questions of test content, length, difficulty, scoring. All the important decisions about the purpose of the test and the educational objectives which it covers must be made at this initial stage;
- (2) writing the test items: subject matter experts and test technicians often work collaboratively to write items within the framework of the specified values and objectives;
- (3) pretesting the items: gathering and analyzing data on a sample response to the items;
- (4) designing preliminary test forms;
- (5) pretesting the preliminary test forms: for reliability, difficulty time limits and so forth;
- (6) designing the final test forms;
- (7) administering the test to standardize and validate it;
- (8) preparing a test manual and other materials; and
- (9) printing, publication and copyright.

Test accuracy is of vital importance in test construction. A basic question about tests is: what do they really measure? Good test instruments have three qualities: validity, reliability and usability. The most important quality is validity, the extent to which the test measures what it is supposed to measure. It can be determined by

statistical comparison of the measurements to some outside criterion. If no convenient and measurable criterion is available, expert opinion may be used, which obviously introduces an element of subjectivity.

Content validity, important for achievement tests, is the extent to which the test includes a representative sample of the universe of content and objectives for the area being measured. The author of a national achievement test must attain content validity for varying curricula, complicating the task of achieving a close fit between instructional material and the test. An item by item analysis is necessary, either in designing or adapting an achievement test.

Construct validity is important for aptitude and social/emotional tests. Construct refers to the trait or characteristic being measured, which may be a cognitive aptitude or a personality factor. The construct validity of an instrument reflects its accuracy and effectiveness in assessing the traits selected, and the extent to which the traits give a representative picture of the behavioral or intellectual dimension which the test is trying to assess. The latter should be clearly defined by the test maker, so that the test can be judged in terms of how clearly it measures what the author intends it to measure. A good test is specific and not ambiguous in its intentions. Construct validity is determined by correlation with a known and accepted test, such as the Minnesota Multiphasic Personality Inventory (MMPI), or by expert opinion.

Concurrent validity refers to the actual behavior of students at the time of testing. This may be easy to ascertain because student behavior can be observed. Results of a test for mechanical aptitude can be checked against actual voluntary involvement of students with mechanical activities or hobbies. Follow-up assessment of the same

criterion behavior yields predictive validity, which is essential in tests affecting the future status of people.

It is important to remember that no test is universally valid; it is valid for a particular time and circumstance. Validity must be closely examined in new tests and adapted tests.

Reliability concerns consistency of results; it is attained through minimum error in administration, scoring and interpretation. Usually high validity yields high reliability, but not the reverse. A test can have high reliability and lack validity. Low reliability can result from fatigue or emotional states in students, unfavorable environmental conditions, and other factors.

Reliability is established by means of two types of statistics-- the reliability coefficient and the standard error of measurement. The reliability coefficient is an index of the extent to which the scores are similar to those on a parallel test. The standard error is the extent to which scores vary over a number of parallel or alternate form tests. To compare scores, a retest can be given after a month or so. The alternate method requires two tests of comparable difficulty level with the same objectives and content.

Usability pertains to administrability, scorability, economy, format and comparable norms (to other tests used). Determining usability is in the domain of the policy maker as well as the testing expert.

Some of the processes involved in test construction in a developing country are illustrated by the Ethiopian Oral Amharic Proficiency Test. The Haile Sellassie I University Test Center in Ethiopia has experimented with construction of new instruments as well as adaptation of existing ones. The Oral Amharic Proficiency Test, a test of proficiency in

listening and understanding of spoken Amharic, was developed in the 1967-68 academic year. The items consisted of common sayings, word usages, word meanings, common sense interpretation of statements, etc. Each item has three possible answers: True or False or neither True nor False. Some sample items are: "the best way to get from one side of the road to the other is to take a taxi"; "a window and a door do not mean the same thing"; and "if two individuals are equal in height then they are also equal in weight."

Out of a pool of 500 items, only 127 items were used in the final test. The rest were dropped because of item difficulty, ambiguity, cultural bias, etc., which became evident after repeated administration of the test to different groups.

The correlation between the oral Amharic scores and first and second semester grades ranged from 0.36 to 0.49. If taken as evidence of oral ability for which the test was designed, the above figures are encouragingly high; its purpose was to measure proficiency in comprehension at an elementary level.

The following are some of the problems that were encountered in the development of the oral Amharic test: (1) the problem of sincerity on the part of some of examinees who were used as the trial samples; (2) a reluctant attitude shown by essential faculty reviewers to make "critical" comments; (3) the problem of avoiding culturally biased items; (4) the problem of constructing simple clearcut statements which are either true or false or neither true nor false; and (5) the problem of locating difficult and/or ambiguous items.

Concluding Comments

Objective standardized tests have great potential value for facilitating educational decision-making on a national level. The use of standardized testing programs to provide systematic data on the outcomes of schooling is a relatively new concept, but one that will be increasingly important in system-wide planning. Measuring the effects of instruction is far from an exact science, but it can yield direct information on the quality and efficiency of formal and non-formal educational systems.

The task ahead for planners is to encourage the development of indigenous expertise in test construction and administration by establishing national or regional programs of test development. These programs, already ongoing in some areas, may well prove to be cost-effective. Traditional exams will not be adequate for either planning or selection purposes in the expanding educational systems of developing countries; the limitations of adapted tests will soon be unacceptable to nationally oriented educators.

## CHAPTER X

### MEASURING THE OTHER A's: ATTITUDES AND ATTRIBUTES

We have distinguished the 4 A's of educational outcomes during the years of schooling. In this chapter, we consider attitudes and attributes or characteristics with affective consequences for students. We shall consider attitudes and attributes in developing nations, factors in the school system with noncognitive implications, and some approaches to assessment of two important outcomes: self-esteem and locus of control.

The formulation and assessment of objectives with a bearing on personal and social outcomes presents a new set of problems for the educator. Goals in the areas of attitudes, values and feelings are difficult to define, and it is even more difficult to achieve social consensus on which should be adopted and which should be emphasized. To assess progress toward any goal in this area is technically complex and beset by the same ambiguities and political controversy of definition and assigning priorities.

#### Social Problems and Educational Goals

The ambiguity inherent in analyzing, teaching and evaluating of non-academic skills does not diminish their importance. People in

most countries, when asked what the goals of education should be, will stress outcomes such as loyalty and nationalism, civic participation, moral and character development, appreciation of culture and learning. But these goals, even when explicitly stated, are not translated into quantifiable objectives for teachers and students. Before taking a closer look at these goals in the school system, we should note some of the affective implications of existing institutions and traditions in society as a whole.

Many of the major social problems in developing countries are caused or exacerbated by social-psychological traditions with ingrained attitudes and inhibitive value systems which are antithetical to national social and economic development. National leaders look to education as a prime force in the complex task of dismantling regressive customs and encouraging new mores, while at the same time maintaining and strengthening the vital cultural and social traditions. Some typical social problems in developing countries with affective implications are:

Rigid social class stratification: economically and racially defined social class systems, such as those derived from feudal traditions in Latin America, result in occupational channeling and are an obstacle to communication and cultural integration.

Fatalism: the religious and philosophical values common in South Asia of renunciation of worldly ambition and of passive acceptance of whatever life brings act as an effective brake on social and technological change.

Alienation from the indigenous culture: long periods of colonization in many parts of the world, particularly Africa, tended to fragment and denigrate existing cultural values and patterns of life. People in urban areas may have adopted European practices and values (for example, in their educational aspirations), which are not rewarding or functional for the people--as individuals or as a society. "Africanization" expresses the desire of African countries to nationalize institutions and initiate a cultural renaissance.

Atomization, or lack of societal unity: primary loyalties to kinship, caste or tribe compete with and tend to undermine national efforts toward unification and development. Urbanization in Middle Africa is one force for detribalization and acculturation of new values; but can have negative consequences as well--alienation and loss of a sense of identity. In South Asia the sense of national unity is obscured by the orientation of people toward family and caste. Kinship orientation reduces mobility which along with fatalistic value systems tends to sharply limit individual striving and entrepreneurial endeavors. Furthermore, lack of faith and trust in others outside the immediate family or tribe lowers the level of interpersonal cooperation in government and business.

Depressed status of women: strictly delineated roles for women, such as Pakistani purdah requiring seclusion and

non-participation in civic affairs, reduces the developmental and democratic potential of a society.

Social changes are occurring in developing nations, however, at a rapid rate; the roles of many people are likely to be drastically altered, whether they are prepared for change or not. The process of modernization involves, as we indicate in Chapter VIII, not only institutional changes in society, but psychological changes as well. Modern societies demand rationalism, adaptability, initiative and a range of other attributes.

Encouraging these abilities, which involve a fusion of cognitive and noncognitive skills, is part of the task of educators. Another part of the task is to correct those aspects of schooling which may have damaging effects on the child's self-identity in the framework of his or her culture and nation.

Leaders in developing nations recognize the attitudinal consequences of education. Goals framed by education ministers in many regional conferences in the last few decades reflect a broad range of national interests, including noncognitive outcomes. For example, a major recommendation of the Commission on Primary Education for the Conference on Education and Scientific and Technical Training in Nairobi (1968) was that African governments should "ensure that the primary education systems of African countries... (b) contribute to the strengthening of national unity; (c) bring about the social and cultural integration of children in the community; and (d) act as factors of change and of economic and social development."

Policy statements from this and other conferences show sensitivity

to the detrimental effects of uprooting children from their families and cultures, often a consequence of secondary education. In Learning To Be, Faure notes that the schools "inculcate values into school-children which estrange them from their surrounding, feeding intellectual and material ambitions which are becoming harder and harder to realize in a rural setting. Schools thereby push young people out toward the towns having failed to instill in them the kind of values which should make people attached to their everyday surroundings."<sup>1/</sup>

A balance must be struck between retaining vital cultural well-springs and cultivating new patterns. These concerns are not contradictory, if nations can learn to absorb changes without creating new but equally rigid institutions. An Arab educator states: "We therefore have to see that our education produces citizens who are well equipped in science and technology who are resourceful, adaptable, adaptive, responsible, realistic, efficient, enthusiastic, effective, and conscious of their problems and the way out of them. The concepts, the attitudes, the skills and the modes of behaving acquired in the past have to undergo changes for the present and more profound changes for the future."<sup>2/</sup>

#### Implications of Schooling for Personal and Social Growth

As indicated earlier, education in traditional societies was functional. Many nonliterate societies have formal learning structures

---

<sup>1/</sup> Faure, E., et al., Learning To Be (Harrap: UNESCO 1972).

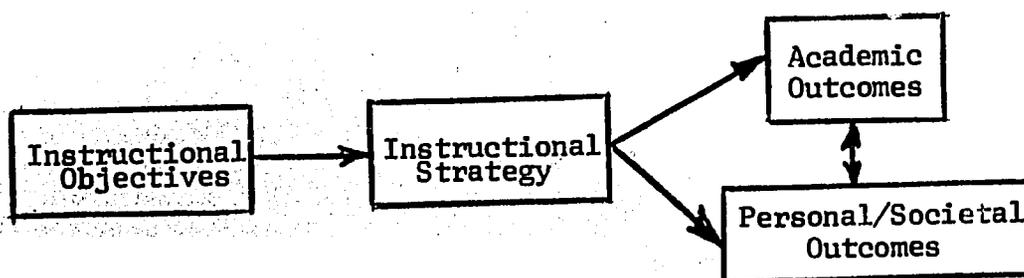
<sup>2/</sup> Abdel, Aziz Hamid El-Koussy, "For a Self-Criticism of Education in the Arab Countries," Prospects: Quarterly Review of Education UNESCO, Vol. III, No. 1, p. 65, Spring 1973.

based on observation, in which the young master large bodies of information and skills in military training, music, geography, religion, and language. In addition, children learn skills connected with their future roles--farming, hunting, housekeeping. A common characteristic of nonliterate education, in the past and today, is that the process of learning is emotionally charged and personalized. The knowledge of the society is highly valued and is transmitted individually or in small groups by teachers with high social status.

In organized and sophisticated early societies, education became more formalized but no less concerned with personal development. Early schools in one nation, for example, emphasized moral and religious education, calisthenics, and ethics. Gradually, as schooling became more institutionalized, the course of study became progressively less congruent with the skills and knowledge learned informally in daily life, more abstract and verbal, and less concerned with the moral and emotional lives of children. In many parts of the world, the schools represented an oppressive foreign culture and actively denigrated the culture of the children, very likely with disastrous results for the individual's self-perception. (Such results are reported by studies of the self-concepts of minority group children in the United States.)

It should be clear, then, that outcome assessment for planning and reform must treat many aspects of the lives of children in schools. Curriculum, teaching techniques, administrative patterns and other factors in formal and nonformal educational structures have important effects on student development in nonacademic areas. Some of these

effects will be described in later chapters in connection with the feedback effects of outcome measurement for curriculum, teaching and administration. It is important to note the reciprocal influences between academic and other outcomes. Attitudes and values effect learning, and are altered by it. The diagram below represents an instructional paradigm in which instructional objectives are phrased (in measurable terms), an instructional strategy devised for achievement of the objectives, and the objectives achieved are measured.



#### Attitudes and Attributes as Inputs

Attitudes and attributes function both as inputs and outputs of education. A child's attitude toward mathematics may greatly influence learning and performance in that area. At the same time, success or failure in mathematics will influence the child's self-concept. In the past, more attention has been directed at personal-social attributes as inputs. Much research in educational psychology has been oriented toward discovering relationships or correlations between personality variable and achievement, with the idea that psychological tests might be used as predictors of academic success. Some of the personality factors that have been found to have a positive relationship to academic achievement include independence, impulse control, achievement motivation, ability to make consistent judgments,

persistence, order, endurance and stability. However, the research results have not been entirely consistent with respect to the achievement implications of most personality variables. Many studies have found no relationship or a weak relationship between achievement and general adjustment as measured by the major psychological batteries such as the MMPI (Minnesota Multiphasic Personality Inventory). Yet we cannot conclude personality factors are not crucial to learning, because the wealth of research studies indicate that they are. We simply cannot point to particular variables as the causal factors.

There are several possible explanations for the lack of clear-cut findings. One is that personality factors interact in ways we do not wholly understand, and, therefore, studying characteristics in isolation is likely to be misleading. Even where multivariate, or multiple factor, studies were employed, relationships tended to be weak (and one must remember that a relationship or correlation between two variables does not mean a cause and effect relation exists). The studies often implicitly assume that the individual operates in a social vacuum; it may be that the social setting of the performance is an important variable.

Another problem is that many of the instruments used in this research were developed by clinical psychologists for the diagnostic identification of maladjusted individuals. These tests tend to be slanted toward negative personality traits or neurotic disorders; they often are quite inappropriate for educational assessment. And since what constitutes "maladjustment" is culturally determined, developing

countries are particularly cautious about using foreign psychological tests.

In the last few decades educational psychologists have developed many instruments to measure particular attributes in a school setting. However, their research orientation makes them impractical for broader assessments because of basic orientation, dated or inappropriate items, impractical mechanics, and so forth. Published tests and others discussed in the literature were often designed by investigators with limited financial support and, therefore, have not been subjected to the extensive and rigorous procedures which produce reliable achievement and aptitude tests. In developing nations, there is the additional problem of translation and cultural adaptation of published instruments, which are more complex than in the case of a cognitive test.

#### Assessing Attitudes and Attributes as Outcomes

Many educators have devoted much attention to the importance of the affective domain. Benjamin Bloom and two associates designed in 1964 a classification model for human behavioral outcomes in the affective domain.<sup>3/</sup> Choosing internalization as the ordering principle for the universe of possible human response, they described a hierarchy of five broad categories of response to stimuli:

level 1.0 Receiving (the learner is conscious of a phenomenon or object and is inclined to pay attention to it).

---

<sup>3/</sup> Krathwohl, D.R., et al., Taxonomy of Educational Objectives Handbook II: The Affective Domain, David McKay Co., N.Y. 1964.

- 2.0 Responding (the learner actively and voluntarily attends to or complies with a phenomenon).
- 3.0 Valuing (the learner believes that a thing, phenomenon or behavior has worth).
- 4.0 Organization (the learner begins to build an organized value system).
- 5.0 Characterization by a Value or Value Complex (the individual acts and responds in terms of his/her system of attitudes and values, or philosophy of life).

The authors have described each category and the sub-categories in each by including sample educational objectives illustrating development on that level.

Other educators have concentrated on developing programs or intervention strategies to further personal and social growth by students. Some techniques include group and individual counseling, play therapy, and dramatic activities, reinforcement approaches in a group setting, curriculum changes (for example, to focus on material students can identify with) or methodological changes (for example, structuring the instruction so that there is a high probability of success for the student).

However, despite some promising programs, assessment of affective student outcomes of education has been largely neglected. Noncognitive objectives are rarely included in instructional programs: as a result, their assessment as desired outcomes is at best an unsystematic

- 11 -

enterprise, often limited to evaluations of scattered special programs. Ordinarily, in the process of teaching even the affective objectives which have been explicitly stated tend to fall by the wayside and organized assessment is forgotten. Even when teachers are consciously aware of their affective objectives for their students, they often feel that they can best evaluate these outcomes subjectively--by observing their students' behavior. Another factor, in many nations, is the reluctance of educators to evaluate personality factors in students. In the last few decades it has been assumed by many that noncognitive development was an automatic byproduct of cognitive growth, though this has not been convincingly demonstrated. There is great interest at present in specifying nonacademic goals and tracing non-academic development, but this interest is stymied by the lack of standardized instruments.

An educator wishing to implement outcome assessment of attitudes and attributes faces a twofold problem. First, how is one to choose the significant variables or a particular variable from the vast array of psychological and social attributes that have been identified and measured? Secondly, how is one to locate good instruments to measure these variables? The Chart on page 40 represents a framework or analytical structure of potentially relevant personal attitudes and attributes. This general classification scheme permits differentiation between broad types of characteristics.

Personal-social characteristics are particular personality traits which are specific individual expressions of the polarities which form

the general dimensions of personality. Controlling mechanisms are the individual's style or manner of dealing with or organizing phenomena, and represent a fusion of cognitive abilities and personality factors. These characteristics are often called "cognitive styles" or "cognitive controls." An example is the characteristic called field dependence, which refers to a person's ability to differentiate an object from its background. A test for field dependence is the rod and frame test, which measures a subject's ability to hold a rod perpendicular to a frame which is not perpendicular to the ground. Performance on this task is seen to predict the individual's approach to a broad range of tasks--it indicates whether phenomena are experienced as discrete from the field or background in which they are contained, or whether the subject responds globally to the dominant organization of the situation.

The third category includes values (beliefs or commitments; general attitudes that are fully internalized), attitudes (responses made to socially significant elements in the individual's environment) and interests (the manifestation of an attitude in an individual's activities). While the general dimensions of personality are universally shared, the expressions of those dimensions are in part culturally determined; the most concrete category of interests may be quite localized (interests reflecting local institutions geography and climate, and cultural activities).

Educational researchers generally investigate the categories of personal-social characteristics and controlling mechanisms. On

the classroom level, teachers are more concerned with the third category of values, attitudes, and interests, which are the focus of the Bloom taxonomy discussed earlier.

On the level of system-wide assessment of educational outcomes, the significant noncognitive factors in all three categories listed here should be identified, using criteria derived from the educational goals of the nation. A group of characteristics felt to be important in national development (for example, autonomy/independence, locus of control, conceptual style, coping style, self-esteem, attitudes toward community and nation) could be selected for the development of assessment measures to be applied longitudinally. These standardized test instruments, applied in standardized testing programs, will help determine whether the school system is reaching its goals. Some aspects of this process are treated in the next sections.

### Selection of Instruments

We focus here on "pad and pencil" tests as the most efficient and economical method for wide scale measurement of the qualities of personal development judged important to a nation. Other effective techniques such as observation, individual and group interview, or projective techniques, may be useful on a classroom or program level.

Most existing paper and pencil tests use the self-description mode of assessment. Description by students of their past and current experience in a structured format is the most fruitful method for systematic group assessment of psychological characteristics, cognitive controls, and attitudes, values, and interests. This category includes

devices such as questionnaires, checklists, self-ratings, biographical inventories and attitude scales. As noted before, very few existing tests are directly applicable to school situations in developing nations. However, they are useful as examples of measures which are needed. Although they lack validating support, are sometimes difficult to acquire, and are ambiguous in interpretation, these existing instruments are the prototypes for the tests that should be developed for use in schools. Both developed and developing countries must apply resources to designing acceptable instruments for measuring these nonacademic outcomes, which are univerrally seen as important.

Several guides have been published in the United States which describe and evaluate published tests in the personal-social development area. These include:

Buros' Mental Measurements Yearbook: The most comprehensive source of test information, the Yearbook includes availability and administration information, a list of published references to the instrument and often a detailed critical review. The seventh edition, published in 1972 contained listings for 1,157 tests and reviews of 798.

Robinson - ISR Series: Robinson and colleagues at the Institute for Social Research, University of Michigan, have published three volumes of test descriptions, Measures of Political Attitudes, Measures of Occupational Attitudes, and Measures of Social Psychological Attitudes. The last was revised in

- 15 -

1973 and includes descriptions and evaluations, with sample items, of 126 scales and measures.

CSE Test Evaluations: The Center for the Study of Evaluation of the University of California has published several books reviewing tests for school children according to detailed criteria. Two of the volumes are: Elementary School Test Evaluations (1970) and Preschool/Kindergarten Test Evaluations (1971).

Other sources for information are test publisher's catalogues, the quarterly Test Collection Bulletin of the Educational Testing Service in Princeton, New Jersey, measurement textbooks abstracts and indexes such as Psychological Abstracts, journals and periodicals relating to educational measurement, and doctoral dissertations.

#### Measures Used in Outcome Studies at Haile Selassie I University

As examples of personal-social measures now available we will look at a few of the scales used by the Haile Selassie Testing Center, and then examine two variables--self-concepts and locus of control--in detail. Some considerations in adapting measures of attitudes and attributes will be considered later.

Personal-social measures must be selected in the light of the particular culture in which the measures are to be used, and the particular outcomes derived. Using criteria of availability, ease of modifiability, and potential for illuminating hypothesized relationships between community and school input variables and certain educational outcomes, and for the prediction of university performance, the following measures were seen as potentially useful. These

characteristics are also related to qualities implied by modernization (see Chapter VIII). The measures were modified for use with graduating high school seniors. A description of the scales and, sample items are given below (all items require an agree or disagree response).

1. Anomie. A state of cultural and personal disorganization in which the individual is unable to refer his behavior and that of his associates to any stable set of standards. The individual caught in such circumstances is said to respond by developing the psychological state of alienation. Measures of anomie have not typically been included in educational outcome assessments. However, such measures could be included, since one educational goal of developing as well as developed countries is in reducing alienation and increasing societal participation.

Sample items: It's hardly fair to bring children into the world the way things look for the future.

These days a person doesn't really know on whom he can count.

2. Status concern. A measure of attitudes toward status and mobility. That is, the value placed by the individual on symbols of status and on the attainment of higher status. Low responses on this measure would indicate achievement of a desired educational outcome of nonconformity. The scale consists of 17 items.

Sample items: The raising of one's social position is one of the most important goals in life.

Before joining any civic or political organization it is usually important to find out whether it has the backing of people who have achieved a respected social position.

3. Test anxiety. A measure of anxiety created by having to take tests. There is evidence from some research that there is typically a negative correlation between high test anxiety and test performance, although a moderate level of anxiety may facilitate performance. Given the importance attached to tests and test performance in Ethiopia, test anxiety is an important construct. Studies related to educational outcome assessment therefore must consider the impact of test anxiety on test performance.

The studies conducted in the University Testing Center using a measure of test anxiety are primarily methodological. They are designed to seek information on the extent to which test anxiety influences performance on various kinds of tests and measures and on the demographic correlates of test anxiety.

Sample items: If examinations could be done away with, I think I would actually learn more.

Thoughts of doing poorly interfere with my performance on tests.

4. Social desirability. This is a methodological scale which is used to study the validity of the various test and scale items. The Social Desirability Scale measures the respondents' tendency to describe themselves in socially desirable terms. In measurement studies, correlations between test item and social desirability scores should be low or negligible. If they are not, it is not possible to separate responses to the test item from the respondents' tendency to place themselves in a favorable light. Items for which this is the case must be eliminated from any test or scale. The measure of social desirability modified from use at HSIU was that

developed by Crowne and Marlowe (1964) and consists of 28 items.

The response set--the tendency to respond to a test item in a manner which is independent of item content--contributes to the invalidity of test items. The Crowne-Marlowe Scale provided a measure for control of social desirability response sets. A second type of set contributing to the invalidity of item response is the tendency to agree to a test or scale item independent of content (acquiescent or agreement response set). This tendency is controlled by wording scale items so that an approximately equal number are phrased in the agree and disagree directions.

Sample items: I have never intensely disliked anyone.

No matter who I'm talking to, I'm always a good listener.

### Self Concept and Locus of Control

These two constructs, which encompass the individual's view of himself and his view of the world in relation to him, are seen by many investigators as central variables. We will examine the constructs and some assessment instruments for each.

#### I. Self Concept

The self concept (one's view of oneself) is a central feature in the personal world of every individual. As a subject for study and assessment, it is complex but has the advantage of cutting across a number of other variables such as motivation, needs, values, and attitudes. Human beings attach meanings and values to things, people, and events; the values attached to the self evolve in the course of

- 19 -

interaction with the external world and in turn guide the behavior of the individual. Self concept is a dynamic personality trait, changing with growth and experience. A practically identical trait often measured is self esteem, which means liking and respect for oneself, or personal judgment of worthiness. A highly related construct (personality trait) is self acceptance.

The importance of self esteem comes from its pervasive effect on behavior. The way people feel about themselves colors their perceptions of the world and their responses to it. Esteem attributes are more subject to perceptual distortion than physical attributes like height or wearing glasses. And since self concept is private and invisible, it may lead to misinterpretations of motivations and intentions by others. For example, if a boy feels that he is a bad reader, he will see reading aloud as a threatening and painful experience and will avoid it. A teacher might interpret this as a "bad attitude" or laziness. The teacher's reaction may further lower the student's concept of himself as a student and reader.

Dimensions of Self Concept: Throughout this century, behavioral scientists have produced numerous research studies and developmental theories concerned with self concept. Some of the internal dimensions, or subselves, that have been described are the identity self (one's basic perception of self identify, which influences and is influenced by one's behavior); the behavioral self (self image as expressed through behavior); the judging self (the self functioning as observer, standard setter and evaluator);

and the ideal self (one's image of how one would like to be).

There are external dimensions as well--self concept varies according to frame of reference. One might have different images of the self as student, citizen, worker, or family member. Self concept as a whole is a composite of the esteem attached to all the subselves. A student may have a low concept of himself as a student but high self esteem as a family member. Everyone's behavior varies from one situation to another, but a well integrated person will have a consistent feeling of worth and self regard while realizing that his abilities in different areas vary. Self theorists see self-actualized people as those who are inner-directed, motivated by personal goals and not the need for approval of others, and consequently lead rich and effective lives.

Results of Research: Self esteem research has been complicated by the fact that it is difficult to correlate the data from the hundreds of studies that have been done. Researchers have tended to design their own tests rather than use existing ones, and have focused on specialized problems so that broad conclusions are difficult to formulate. Some results of research follow.

Many studies have examined the discrepancy between the self image and the "ideal" self image of a person. It was originally hypothesized that a high disparity indicated maladjustment, but U.S. studies have shown that the disparity increases with age, IQ, and cognitive complexity, among normal well-adjusted populations.

Evidently older and more intelligent people in the United States have higher expectations of themselves and better defined perceptions of goals. Cross cultural studies have indicated the influence of the cultural environment on this discrepancy. One study comparing American and Indian 12-year-olds, both with high socioeconomic status, found a much higher congruence between real and ideal self concepts among the Indian children, particularly the Indian girls.<sup>4/</sup>

It has been found that a generalized high self concept has a positive but weak relation to intelligence, and a stronger positive relationship to specific self concept of school ability and self report of school grades.<sup>5/</sup> The importance of the opinions of "significant others" to self concept has often been noted. Investigators have found a high correlation between the child's self concept and the teachers' reports of perceptions of them, and also between the children's self perceptions and their parents' perceptions of them. Other studies indicate that self esteem is related to sex role identity for adolescents.

Researchers have also found a positive relation between self esteem and academic success, but it is not a simple one-to-one

---

<sup>4/</sup> Swart, M.S. and Swart, R.C., "Self-Esteem and Social-Personal Orientation of Indian 12 and 18 Year Olds," Psychological Reports 1970, pp. 27, 107-115.

<sup>5/</sup> Bachman, G., Youth in Transition, Vol. II: The Impact of Family Background and Intelligence of Tenth Grade Boys, Ann Arbor, Michigan, Institute for Social Research, 1970.

correspondence, since many other variables intrude. Coopersmith found that a high discrepancy between goals and performance, or a high discrepancy between self and teacher evaluations, led to repeated failure for 5th and 6th graders.<sup>6/</sup>

Another general finding is that self esteem is affected by the experiences of the individual. In American studies, candidates for office who were not elected, and university students carrying heavy work loads, lost self esteem.<sup>7/</sup> On the other hand, one study found that the self esteem of black students increased after outstanding black speakers were brought to the school during a school year.

The self concept is formed in the context of a particular cultural milieu and a particular peer group and it may be relatively independent of social and economic status. Some studies in the U.S. demonstrated that disadvantaged black children (defined as those coming from low income families living in low rent or subsidized housing) had higher self perceptions than advantaged children.<sup>8/</sup>

---

<sup>6/</sup> Coopersmith, Stanley, "Self Esteem as a Determinant of Selective Recall and Repetition," (Doctoral Thesis, Cornell University), 1958, pp. 130-131.

<sup>7/</sup> Ziller, R. C. and Ziller, L. H., "Political Personality," Proceedings, 77th Annual Convention of the American Psychological Association, 1969, p. 442.

<sup>8/</sup> Soares, L. M. and Soares, A. T., "Interpersonal Perception of Disadvantaged Children," Proceedings, 79th Annual Convention of the American Psychological Association, 1971.

Measurement of Self Esteem: Problems abound in the measurement of an attribute which is idiosyncratic and not directly observable. Most self theorists feel that it can be best appraised through self reports. Self report formats that can be used effectively for group assessment are forced choice scales (test takers are presented with two or three statements about something and must choose one) and Likert-type scales (the test takers specify whether they strongly agree, agree, are undecided, disagree or strongly disagree with a statement). Another format is the adjective check list in which 300 or so adjectives are listed and the respondent checks the ones felt to be self descriptive.

Some sample self esteem test items are: (items answered either "like me" or "unlike me.")

"I often wish I were someone else."

"If I have something to say, I usually say it."

"I often get discouraged at what I am doing."

A problem with testing for self concept that has received some attention is that it is difficult to include all the possible important dimensions of self concept in a standardized test. A test may contain items about school, and other situations without including any reference to areas from which some individuals may derive high self esteem--for example, unusual hobbies or activities. Some self concept scales are generalized while other are specific to the school situation. Educators must decide if they are interested primarily in the students' academic self esteem or

if they need more generalized information.

## II. Locus of Control

In recent years, an increasingly productive area in psychological research and measurement has been the investigation of personal cognitive styles of controlling mechanisms. Cognitive style has been defined by the psychologist Jerome Kagan as the "stable individual preferences in modes of perceptual organization and conceptual categorization of the external environment."

A particular cognitive control characteristic which has been identified as important to individual development and on a broader scale to national development is locus of control.

The concept behind the construct called locus of control is that people view events which affect them as resulting from their own behavior or as controlled by outside forces: fate, chance, or powers beyond one's control or understanding. People who tend to see a cause and effect relationship between their own actions and things that happen to them are called "internal"; they feel that some control lies in themselves. People who believe that most of the things that happen to them are caused by external forces are called "external."

Many behavioral scientists believe that this characteristic affects a great deal of a person's life, as well as the character of a society. A feeling of control over one's environment probably facilitates change, mobility and growth within that environment. A feeling of lack

of control may lead to acceptance of the status quo and hopelessness about the possibility of changing it. The amount of effort invested in any activity is a function not only of motivation, but also expectancy of success; both of these elements are tapped in locus of control measurement. Locus of control is sometimes referred to as IE, an abbreviation for "internal-external" control.

Like self concept, locus of control is a complex attribute. One must be cautious about rigid interpretations or values to particular outcomes. An individual's outlook on life must be tempered by reality. A person who attributes all responsibility to himself would be deeply frustrated in a situation that really is beyond his control: for example, being confined as a prisoner of war. In such a situation, it would be more sensible to have an external orientation. Similarly, it is risky to make comparisons of locus of control attributes across cultures or between different socioeconomic or ethnic groups. The realities of power vary widely between societies.

However, locus of control is a meaningful dimension of a student's behavior and response to school and adaptability in later life. It is probably in the interest of most nations to promote an internal orientation among its citizens for social and economic development.

Research: Most research indicates that people are hindered by external locus of control beliefs. Achievement is depressed, health and safety precautions are less likely to be taken, delayed rewards are less valued, conformity and docility are higher. High achievement test scores were found to be positively

related to internal locus of control. In a large scale project, composite achievement reading, math and language grade placement, and academic self concept were positively related to internal control for boys.<sup>9/</sup> The positive correlation between internal control and achievement has also been found for girls who scored low on social desirability scales, in other words, who were not dependent on others for approval.<sup>10/</sup> In studies of adults, internal workers have shown improved job performance, and more highly developed personal qualities relating to employability and job success, in contrast to external workers.

Measurement: Some of the dimensions of locus of control that have been identified by factor analyses of pools of items from locus of control scales have been labeled: (1) control ideology or the subject's belief about the extent to which people have control generally, (2) personal control, or the extent to which the subject believes in personal control, and (3) system modifiability, referring to control over world affairs, war, racial discrimination, etc. Another aspect of the construct is the application of a positive or negative value by the individual

---

<sup>9/</sup> Martin, F., The Creative Organization of Positive Experiences, Research project Gwinnett County Schools, Lawrenceville, Georgia, 1972.

<sup>10/</sup> Norwichi, S. and Walker, C., "Achievement in Relation to Locus of Control: Identification of a New Source of Variance," Journal of Genetic Psychology (in press).

to his orientation--one might view the world as controlling and malevolent or controlling and benevolent. Another question that has been raised concerns linking fate, chance and powerful others together as external forces. One researcher has devised a scale which separated powerful people and chance in scoring for external control.<sup>11/</sup> A further distinction that has been made relates to differentiating between control over positive and negative events. Some locus of control tests are designed to yield both scores. There is a great difference between a student who accepts responsibility for success and denies responsibility for failure and a student who follows the opposite pattern.

Despite the complexity of the construct, researchers in the United States have had considerable success measuring IE. Some are attempting to develop unidimensional scales by omitting situation-specific items to arrive at a single overall score. Some are specific to control over and responsibility for academic success and failure. Other social scientists prefer multidimensional scales.

The formats of paper and pencil tests include forced choice between statements, the yes-no response to questions, and Likert-

---

<sup>11/</sup>

Levenson, H., "Distinctions Within the Concept of Internal-External Control: The development of a New Scale," paper presented at the American Psychological Association, Hawaii, 1972.

type format. Several non verbal formats with cartoon pictures have been developed for children. These must be read aloud by the test administrator. Some examples of a locus of control item are:

"When you do well on a test at school, is it more likely to be

A. Because you studied for it, or

B. Because the test was especially easy?"

"Do you believe that most children are just born good at sports?"

"Do you think it is better to be smart than to be lucky?"

The significance of the self concept and locus of control constructs must be analyzed in terms of the present cultural milieu of a society as well as its future needs and goals. There are many unanswered questions about the meaning of personality variables in different societies. For example, what is the relationship between self concept and locus of control within different status groups? One hypothesis is that in a highly traditional society members of lower classes may have a high level of self esteem (derived from the security of a class structure with clearly defined roles and expectations for individuals) combined with an external locus of control orientation. This configuration might be reversed from upper class members, who have greater mobility but higher self expectations. Another problem to be more fully investigated is the effect of modernization, and the requirements of modernization, in relation to these variables. Generally it can be said that high self esteem and internal locus of

control are desirable outcomes; but much remains to be learned about the dynamics of these characteristics for different groups (i.e., urban workers, farmers, women, entrepreneurs) within the developing context.

#### Test Adaptation and Development: The Ethiopian Experience

The Haile Sallassie I University Testing Center has experimented for several years with construction and modification of nonacademic measures for use in Ethiopia.

The adaptation of instruments developed in other countries is less than ideal. However, it should be recognized that there are many commonalities in human behavior that cut across cultures. To develop new instruments for each situation in which the measures are to be applied is not always possible or necessary. It should not, however, be assumed that tests and scales can be taken from other cultures uncritically. The measures must be subjected to the critical appraisal of individuals knowledgeable about the culture and modified as appropriate. Moreover, the measures must be subject to psychometric study to determine their reliability and their utility for the particular assessment program in which they are to be used. Only when these steps have been undertaken and the results known, can an instrument be said to be satisfactory for use in settings other than those for which it was developed.

Modification of tests and measures developed in the United States took the form of elimination of items seen to be culturally inappropriate

- 30 -

or rephrasing of questions to eliminate inappropriate words or phrases. A third possibility--the addition of items of similar content but designed to reflect the local situation--may be used in future studies.

Decisions about modifications of scale items are judgmental. No detailed rules were developed; however, they concluded that those who engage in test development and modification should have a thorough understanding of the constructs underlying the test or measure to be modified, and should be well informed about the culture in which the scale is to be used. An Ethiopian educational psychologist and an Ethiopian cultural anthropologist screened and modified the scales used in studies conducted by the Testing Center. Each test item was reviewed by the judges for its clarity and relevance to Ethiopian culture. The judges eliminated some items, and suggested specific changes for others.

Item Appraisal in Operation: Below are given examples of items judged to be questionable or inappropriate within the Ethiopian context, and which were eliminated or modified at the suggestion of one or both experts. As already noted, the process is of necessity subjective and will vary from one culture to another, as well as with the experience and orientation of the judges.

#### Items Eliminated

The average citizen can have an influence on government decisions.

This world is run by a few people in power, and there is not much the little guy can do about it.

(Internal - External Control)

With enough effort we can wipe out political corruption.

It is difficult for people to have much control over the things politicians do in office. (Internal - External Control)

Before voting I thoroughly investigate the qualifications of all the candidates. (Social Desirability)

My table manners at home are as good as when I eat in a restaurant. (Social Desirability)

If I could get in a movie without paying for it and be sure I was not seen, I would probably do it.

(Social Desirability)

Within the context in which these test questions were administered (i.e., as part of the Aptitude Test battery for university admission given in all provinces of the Empire) political questions were deemed inappropriate, lest it be perceived that student responses to such questions would influence university admission. However, a byproduct of removal of certain political questions from the Internal-External Scale improves its unidimensionality. Certain data (Gurin, 1969) indicate that the scale has several components, among them a political one. When the political items are removed, as in our use of the scale, the factorial structure of the instrument was improved.

The social desirability items described above were eliminated because of their unsuitability for utilization throughout the

Empire. For example, the question concerning getting into a movie without paying would be appropriate for use in Addis Ababa where there are many movie houses, but inappropriate in many other communities where there are none. The same rationale was applied to the question concerning eating in a restaurant. In addition to the obvious class bias associated with eating in a restaurant, restaurants are more likely to be available in the larger population centers than in the small communities, thus rendering it an unsuitable test item.

#### Items Modified

I always try to practice what I preach. (Original Social Desirability item)

I always try to do what I tell other to do. (Modified item)

Although people sometimes compliment me, I feel that I do not really deserve the compliment. (Original Self Concept item)

Although people sometimes praise me, I feel that I do not really deserve the praise. (Modified item).

I become panicky when I think of something I have done wrong (or might do wrong in the future). (Original Self Concept item)

I become scared when I think of something I have done wrong (or might do wrong in the future). (Modified item)

As can be seen, modified items were those perceived to require only a minor change in wording, the central thought of the message being retained. For example, in the first modified item above, the idiom "practice what I preach" was one which was perceived as not likely to be uniformly understood by Ethiopian youths and hence was replaced with the phrase "do what I tell others to do," preserving, it is assumed, the original meaning of the item. In the two items following, only a single word has been changed, these because it seemed that they would be less likely to be understood than would the words which replaced them-- "praise" for "compliment" in one item, and "scared" for "panicky" in the second.

#### Test Design: Impact of Language on Outcome Measurement

Ethiopia is a country of several linguistic groups, of which Amharic is the official language and English the second official language. In upper elementary and secondary schools, and in the university, English is the language of instruction and in most instances is the language used in test construction. However, the effect on test performance of testing in one language or another is not entirely clear. For example, we need answers to questions such as the following: How does the language in which the test question is phrased affect the test's validity and reliability? From the point of view of the test constructor the central problem is to devise the most accurate measure of a trait which has been identified as important for some purpose. In devising such measures the language

in which the test question is phrased should have no influence on the candidate's score. While this situation would be ideal, presently available evidence indicates that this is not the case. We will review the work of the Testing Center on the problem of language and its implications for personal-social and cognitive style testing.

In one study conducted by the University Testing Center, identical tests of geography and intelligence were constructed in Amharic and in English and administered to random samples of 474 candidates for a government clerical position. The results were indeed revealing. Mean scores on the geography test for Amharic and English forms were virtually identical with, respectively, average scores of 20.84 and 20.88. However, the case was entirely different for the test of intelligence. On this measure higher mean scores were registered for the test constructed in Amharic where the mean score was 29.47, as contrasted with the mean score of 21.82 for the group tested in English. These results were indeed unexpected and bear discussion even though any explanation offered must be ad hoc. One possible explanation resides in the nature of the task. The geography test was relatively straightforward and required little sophistication in language. The ability test, on the other hand, required active manipulation of language, and it seems likely that the respondents had greater skill in manipulating Amharic than in manipulating English.

In the light of the above findings it seems reasonable to ask if there are suggestions for use of tests which are constructed in one language or the other. A decision on whether to construct a test in

one language or the other would seem to depend on the use to which the test results are to be put. If one is interested in measuring the intellectual power of an examinee, for example, then that method which is most likely to unequivocally reflect such power should be chosen. On the basis of present limited data, the Amharic Test should be chosen. On the other hand, there may be instances in which knowledge of English is important and hence a test constructed in English would be chosen.

The above "rules of thumb" are merely suggestive, since additional questions need to be asked before constructing a test in one language or another. One question relates to test reliability. In the studies conducted at the Testing Center, higher reliability is generally achieved in tests constructed in English. However, this finding must be viewed tentatively since the tests were first constructed in English and translated into Amharic. The fit between the languages is less than perfect in many instances, so it is possible that the low reliability of the Amharic tests can be explained by this fact.

A second consideration relates to test validity. It is necessary to determine how well the test predicts some criterion. In this case, whether a test is constructed in English or in Amharic is incidental; the crucial question is how accurately test scores predict the criterion. This is an area needing considerable additional study, but limited data on concurrent validity was revealing. Identical ability tests constructed in English and in Amharic were both correlated with certain other measures of Amharic or English in order to determine how closely

scores on the different measures would be related. The data of this analysis are summarized in Table 1.

Table 1

Concurrent Validity of Tests Constructed in English and Amharic

Test	Reliability (KR-20)	Correlation With	
		Amharic Ability	English Ability
Geography (Amharic)	.645	.450	.400
Geography (English)	.771	.300	.400
DAT-Verbal	.567	.315	.570
DAT-Numerical	.731	.353	.586
Amharic Reading	.821	.442	.431

The one generalization clearly emerging from the data of Table 1 is that tests constructed in the same language correlate more highly with one another than is the case when tests in one language are correlated with those in another (even though tests across languages are ostensibly identical). One implication of these findings is that tests constructed using different languages should not be combined to obtain a total score, since there is some unknown portion of the correlation between measures which is due to language differences.

A third consideration in determining whether a test should be constructed in English or in Amharic relates to how the individual will be ranked on the respective tests. If language bears no relationship to test performance, then a subject's score should be the same, whether the test is phrased in English or in Amharic. Additional

research on this problem should be undertaken.

A fourth consideration relates to preference for taking a test in one language or another--if given a choice. In one study conducted in the Haile Sellassie I University Testing Center we found a strong preference to be tested in the students' own language--in most instances irrespective of the purpose for which the test was to be administered. Seventy-four Haile Sellassie I University students, mostly of freshman or sophomore standing, were given a 20 item "Language Questionnaire" with these instructions:

Below are listed a number of tests which could be constructed in (A) Amharic, (B) English, (C) Galligna, (D) Tigrigna, or (E) Another language, if none of the above is your first or preferred language. For each of the activities below mark the language in which you would prefer to be tested in order to present the most accurate picture of your ability, knowledge, attitudes, aptitude, achievement, or personality.

Items to which the subjects responded were (1) admission to HSIU, (2) admission to graduate school in the United States, (3) assessment of my personality, (4) assessment of my aptitude for music, (5) assessment of my aptitude for a position in the Ethiopian Government, (6) assessment of my achievement at HSIU, (7) admission to graduate school in Germany, (8) assessment of my various interests, (9) assessment of my intelligence, (10) a post in my major field of study, (11) assessment of my vocabulary, (12) assessment of my knowledge of mathematics, (13) assessment of my knowledge of science, (14) assessment of my knowledge of Ethiopian History, (15) admission to undergraduate study in France, (16) admission to undergraduate study in the U.S., (17) assessment of my ability to get along with

people, (18) assessment of my attitudes toward some person, (19) assessment of my attitude toward some institution (such as the Church or the University), (20) assessment of my attitude toward some object (such as automobiles or airplanes).

Of the 20 items listed above, those related to assessment of personality, interests, ability to get along with people, and attitudes towards other individuals, institutions, and objects are nonacademic in nature. The results revealed overwhelmingly that the majority of respondents preferred to be tested in the language of their area. For the six nonacademic purposes listed above 66 percent of those who specified their first language as Amharic checked Amharic as the language in which they preferred to be tested, and 83 percent of those listing Tigrigna as their first language indicated Tigrigna as the language in which they preferred to be tested. For the construction of nonacademic tests in Ethiopia the clear suggestion from these data is that--from the perspective of students at least--the preference is for tests in the students' first language. Among other considerations it seems reasonable to expect that feelings, attitudes, and emotions are most likely to be expressed in the language of the respondent's area (e.g., Amharic, Tigrigna, Guaragina, etc.) rather than in English or some other language, and that there are unique phrases, expressions, and idioms that lose their meaning when translated into another language. Quite often there are expressions which simply cannot be translated successfully into another language. There is good support, therefore, for the view that

personal-social and cognitive style tests are likely to have greater predictive utility if constructed in the respondent's native tongue. However, whether results of tests so constructed are more valid and reliable than those constructed in English or some other language is an empirical question which remains to be answered.

**Chart 1 Cognitive and Personal-Social Style**

**I. Basic Drives, e.g., self assertion,  
fear, gregariousness, succorance, curiosity**

**II. General Dimensions of Personality, e.g.,**

withdrawn vs. involved  
masculine vs. feminine  
rebellious vs. compliant  
expressive vs. restrained

tense vs. relaxed  
sensitive to others vs.  
self centered  
submissive vs. dominant  
active vs. passive

apathetic vs. energetic  
solitary vs. social  
assertive vs. timid  
aimless vs. purposeful

rigid vs. flexible  
happy vs. unhappy  
academically moti-  
vated vs. otherwise  
motivated

**III. Areas of Personality Expression, e.g.,**

**A. Personal-Social Characteristics**

achievement  
affiliation  
aggression  
anxiety  
autonomy/independence  
curiosity  
deference  
dependency  
dominance  
nurturance  
creativity

**B. Controlling Mechanisms**

scanning-focusing  
field articulation  
conceptual differentiation  
tolerance for delay of reward  
internal vs. external locus  
of control  
risk taking strategy  
impulsivity-reflectivity  
conceptual style  
cognitive complexity  
distractability  
coping styles

**C. Values, Attitudes and Interests**

Interests: attitudes toward  
manipulative 1. family  
cognitive 2. groups (political  
aesthetic 3. community  
4. self  
5. tasks (physical)  
6. important be-  
liefs and acts  
7. national and  
international  
8. nature  
9. the future/  
posterity

## CHAPTER XI

### MEASURING OUTCOMES BY TEACHER GRADES AND TEACHER TESTS

This chapter focuses on more limited measures of educational achievement than previously discussed: teacher grades and tests. Standardized or other externally-constructed examinations are usually not suitable for measuring achievement of the objectives of a classroom's specific unit of study or of certain goals of a particular school. They are wholly inadequate for determining whether non-formal educational programs are achieving their objectives. We review the use of teacher tests and other nonstandardized measuring instruments in schools, discuss their purposes and compare them with standardized tests. Then we discuss their use in the developing nations.

In a number of countries, teacher or course grades are the basis for reporting pupil progress to the pupil, parent and school administrators and are used for promotion, graduation and honors. In other countries, including much of the developing world, external examinations are the primary or only measure of achievement in the school. The trend in countries relying heavily on teacher grades as the measurement of school progress has been to supplement them with standardized, external-type tests (the United States, for example). In countries using external examinations almost exclusively, there has been some movement toward

internal assessment of learning (France).

A grade or school "mark" is usually a composite of scores on teacher-made tests, or classroom contribution and of quality of laboratory work, homework assignments and other projects. Thus, it sums up the teacher's evidence of pupil performance in a course for a limited period. Frequently, primary schools in the United States do not assign a single mark for a course. Instead, learning goals are broken down into specific parts and descriptions of various behaviors and performances are used as a kind of checklist.

### Teacher Tests and Their Uses

This manual has been concerned primarily with educational measurement and the data it provides the evaluator who in turn feeds it back through the decisionmaking, planning, and administrative process. But measurement and evaluation are important not only on policy and planning levels; they are an essential part of the instructional process at the classroom level.

In the learning process it is important to measure learning achievement throughout the course of study as well as at the end of a period of instruction. Teacher exams are a means of periodic measurement and have a number of functions:

- (1) to help the teacher evaluate the adequacy of his/her instructional techniques. Just as the fundamental task of educational systems evaluation is to provide information for making decisions on a "high level" or global scale, so teacher-made exams are useful for making decisions at the classroom level about ways of making instruction

more effective. Generally, low scores on an exam, for example, may tell the teacher that the unit was poorly planned or not effectively presented or was introduced to inadequately prepared students--that learning objectives had not been achieved due to ineffective teaching or to lack of effort of students or both;

- (2) to assess student progress. A reliable, valid exam provides information about how the individual student is performing in class. That information is furnished to the student, and he and the teacher may compare his score with previous exam scores to understand his progress;
- (3) to motivate and direct study and learning in the short range. Periodic or frequent quizzes and tests may motivate a student to study and help set goals for future learning. A well-prepared test covers the objectives of a unit of study and helps define and articulate those objectives--and are useful both to the teacher and the student; and
- (4) as a basis for course grade.

#### Other Measures

A teacher's tests do not measure such objectives of the educational system as use of library, skills in committee work, ability to express ideas orally. Nor do they test for curiosity, enjoyment of art, music or literary style, or creativeness. Adequate standardized tests are not available for such important behaviors as independent self-direction, selection and organization of materials for a report, and understanding of other persons.

Some might urge that tests are hardly adequate for determining whether learning has been attained. They produce artificiality of conditions that from some perspectives suggest skills or knowledge that may or may not be evident in behavior. Or behavior may point to skills and knowledge that are not indicated by test results.

Scores on teachers' exams are supplemented, therefore, by other evaluations of student performance. For example, direct observation can be made and recorded of library use, cooperation with others in committee work, public speaking. Also, appraisal of many products of course work can be made--of term papers, essays, outputs of home economics courses and woodworking or mechanical shops, and art work.

It is difficult, when observing or appraising, to select significant and objective criteria, to include broad sampling of behavior or performance, and to remain an objective observer. However, some subjectivity and unreliability may be eliminated if the teacher can determine in advance what to observe and how to observe it. Methods used in observation or appraisal include:

- (1) ranking--define specific criteria to be observed, rank by degree of quality or how well it meets criteria;
- (2) checklist--itemize specific features to be observed; then, perhaps by yes or no or by a check (✓) indicate whether action took place or not and whether it was satisfactory. This method is useful for appraising products; and
- (3) rating scale--observation or impressions over a time period based on perhaps a 5-point number or descriptive word scale (superior, excellent, average, fair, poor). Weaknesses to

be avoided include the "halo" effect (tendency to generalize from one or two attributes) and the tendency to avoid extremes and rate student performance in the middle.

Anecdotal records provide an informal way of noting spontaneous significant or unusual behavior that may contribute background for evaluating achievement. No "judgment" or rating of behavior is involved. Other indications of changes in student behavior may include:

- attendance records;
- assignments completed on time; extra-credit assignments;
- library use: possession of library card; number of books checked out; number reported read (either required or voluntary);
- case histories; autobiographical data;
- extracurricular activities; and
- awards.

Observed behavior patterns may be appropriately used for a number of knowledge and skill situations in developing nations. Perhaps one approach is observation of an experimental type in which the observer records performance after watching the behavior and response of the student observed. Another approach could be reporting on the actual actions taken in the course of daily living arrangements. What a person does represents his behavior pattern, and those actions may be collected and quantified.

#### Differences Between Teacher Tests and Standardized Achievement Tests

Teacher tests and standardized achievement tests usually serve different needs. The teacher prepares his/her own tests as part of the

teaching-learning process for a particular class and to evaluate the achievement of pupils in that class. Standardized achievement tests are given to many students and are useful for comparison of achievement in one school with achievement in another or school norm with national norm. They may not be an appropriate basis for the classroom teachers' evaluation of the achievement of his/her students for they may not measure the particular objectives of the class. A teacher's test is tailored to fit specific needs and is based on the subject matter content and objectives of his/her class; a standardized test is based on subject matter content and objectives common to a large number of schools. Teacher tests are flexible, suited for frequent use on specific topics (although they may cover comprehensive subject matter); a standardized test usually deals with a broad part of knowledge or skill.

If teacher tests use the same criteria as standardized tests, that is, if they are valid, reliable, objective, and easily administered, they are useful indicators of student achievement (Chapter IX). Too often, however, the tests may be hastily or poorly planned, may neglect fundamentals of the course and may sample only minute details. In addition, teachers may rely on essay tests which are quicker and easier to prepare but may fail to adequately measure student progress in mastering course subject matter content. Frequent tests or too-heavy emphasis on "passing the test" may distort the learning objectives, turn the student against teaching (and learning), and generally defeat their purpose.

#### Teacher Grades vs. Standardized Tests

As we have discussed above, teacher grades are based on student performance and are usually a composite of such measures as scores on

teacher-made examinations, student classroom contribution, quality of laboratory and homework assignments. They are often relative, based on comparative class scores. But grading is not necessarily norm-referenced. Criterion-referenced teacher tests and other criterion-based measures could eliminate grading "on the curve." This means that before grading, the teacher should have explicit criteria established. We have also pointed out that the unreliability of grades as measures of achievement may be caused by ineffectual teaching, variability of marking standards of the teacher and/or school, lack of objectivity of teacher evaluation, inconsistency, and so forth.

Standardized tests measure levels of knowledge, skills and competence; they have been empirically developed and the reliability and validity have been checked and norms have been developed. They may measure the same knowledge and skills as teachers' classroom tests, for they usually try to cover only what schools teach and generally indicate whether students have mastered some body of material covered in the test. Norms for the content of these tests are based on typical or composite course materials for a class, so some tests may not cover special objectives of local schools and ability levels of local pupil population. Studies have shown, however, that skills needed to score high on a standardized achievement test and the skills necessary to earn high grades in school are similar but not alike (Jencks).

Teacher grades as a measure of achievement present a number of problems. They are usually influenced by factors besides performance; they may reflect class conduct as well as achievement; some teachers may reward effort as well as ability; different standards are used in different schools or by different teachers in the same school.

Therefore, grades awarded by different schools seldom reflect comparable levels of achievement. In the developing nations, differences between schools and between teachers and their grading could be very great.

The advantages and disadvantages of standardized tests and teacher tests are listed below:

Advantages of standardized tests:

- all participants are evaluated on the same set of criteria;
- tests are objective; and
- standardized tests used at intervals throughout the pupil's school career provide longitudinal information.

Disadvantages of standardized tests:

- achievement measures provide limited information about a student and his skill knowledge;
- tests are given in an artificial situation;
- tests may control instruction and course content;
- may not cover certain objectives of local schools; and
- may not measure knowledge of students who have not studied the usual textbooks.

Advantages of teachers' grades and school records:

- teacher grades are based on cumulative record of performance;
- they show typical behavior and performance in daily classroom situation;
- they provide teacher and student with periodic assessment of achievement; and
- they show more than skill knowledge--may include assessment of behavior that cannot be measured by tests.

**Disadvantages of grades and school records:**

- changes in short-range cognitive and affective changes are recorded and observed by the teacher only during the term or school year, not over the school career; cumulative records may not substitute for longitudinal records built up by standardized tests;
- teacher grades often compare students within a class instead of providing data on how much has been learned;
- subjectivity or bias of teacher: grades may be more influenced by good conduct, docility, neatness than on how much a student knows; and
- pressure from parent or community may influence grades.

A further problem with the use of school grades as a measure of educational achievement is illustrated by problems Ethiopians have noted concerning their practice of using school marks in evaluation of student achievement.

For some time, school marks in Ethiopia have been given some consideration in determining the passing score for Grade 6 candidates. National Examination results count for 70 percent and school results 30 percent of the score. However, the testing department of the Ministry of Education has objected to this combination of results on the basis that school results do not reflect the actual achievement of the candidates. They suspected that the teachers in various schools inflated the school result hoping to upset the National Examination result and thereby increase the number of passers from the given school (Ministry of Education Reports, 1970, p. 29). The argument presented against

using the school results does not appear to be based on sufficient evidence, however, as no information is available to verify whether or not the school teachers have consistently been generous in giving grades to their students over a number of years. If the judgment of the authorities in the testing department of the Ministry of Education is based on a single year's result (which it appears was true in this case), the suggestion to disregard the school results in deciding the passing grade does not seem to be convincing since it lacks sufficient evidence for suspecting the authenticity of the school results. On the contrary, the teachers might be the best judges of the achievements of their students than a test set by a body of examiners who have little or no knowledge of the examinees' ability. One may conjecture that a better selection procedure may be devised if the teachers and the Ministry officials collaborate in the effort to prepare the instrument.

In the meantime, however, it may be helpful to take into consideration the appraisal made both by high school and elementary school teachers when assessing the achievement of a candidate in an external examination. Incorporating the evaluations made by Ethiopian classroom teachers may give a better picture of the general abilities of the students than a result in an examination prepared by an external body which may have very little or no knowledge about the abilities of the candidates.

#### Possible Ways to Adjust Grades

Teacher grades might be used as a supplement to standardized tests in assessing how much learning has taken place, but if grades are to be a useful source of data, adjustments would be needed. Perhaps some set of indicators could be developed to measure school differences. Some

factors considered in such an adjustment are:

- (1) differing school policies on grading;
- (2) rating schools on:
  - (a) number of students going into higher educational institutions;
  - (b) number of former students employed in various job categories (or not employed); and
  - (c) general behavior of school leavers
    - use of library--number who have library cards;
    - subscription to newspapers--read newspapers;
    - use of radio and TV;
    - membership in professional groups or other organizations; and
    - voting record; and
- (3) teacher differences:
  - general qualifications;
  - credentials; and
  - experience.

### Trends in the Developing World

A number of developing countries have been taking a hard look at the examination policy in use in their school systems. Some observers have been critical of overemphasis on passing an exam at different levels.

"The single most important factor affecting the direction of education in E. Africa today is the realization that the vast majority of the children who go through the system will never see the inside of a secondary school, and that they will have to make a living on the land after primary school. An education system dominated by examinations and aimed at preparing primary school

children for secondary school and secondary school pupils for university, cannot meet the needs of the majority of children."<sup>1/</sup>

There are a number of different systems of examinations, admissions and requirements for school-leaving certificates. In a number of countries, a recent survey shows that there seems to be a trend towards a balance between internal and external control of exams and a trend towards a balance between formal exams and school records based on daily work (Atiyeh).

Furthermore, in British Columbia, Iraq, Bulgaria and many school systems in the U.S., students who are high achievers in regular school work are exempt from formal exams. Cumulative school record is a basic consideration.

A policy of selection based on elementary school leaving exams seems to depend primarily on the policy of admission to secondary schools (Atiyeh). Countries with a restrictive admissions policy usually require an elementary school certificate based on an external examination. But in a number of cases, school records provide supplementary evidence (Lebanon, France, Russia, Spain, Libya, Bulgaria). For admission to secondary school in Mali, an elementary school leaving certificate (based on an exam) is required; final selection is made by a national guidance committee after considering candidate's age, scholastic record, aptitude test results.

---

<sup>1/</sup> Kajubi, W. Senteza, "New Directions in Teacher Education in East Africa," International Review of Education, 17, 1971, p. 202.

In the mid 1960's, a UNESCO study suggested that an unnatural emphasis on exams in Asia was an obstacle to improving educational quality. An effort has since been made in a number of countries to improve exam procedures. While this does not mean that most countries have abandoned external exams or placed heavy emphasis on internal or teacher exams, there has been emphasis on using information from exams to provide guidance for teachers, pupils, school administrators and parents and to give instruction in evaluation. In Malaysia, a system of automatic promotion has been adopted and pupil progress is assessed internally by regular tests. Standardized tests are now being constructed within the Ministry of Education. India also has attempted to strengthen internal school assessment to correct over-emphasis on external examinations.

## CHAPTER XIV

### ROSS-NATIONAL EVALUATION OF EDUCATIONAL ACHIEVEMENT:

#### THE IEA STUDY

In 1966 the International Association for the Evaluation of Educational Achievement (IEA) began a six subject, cross-national survey of educational achievement. Some 20 nations, four of them less developed ones, were involved in this evaluation of learning achievement in science, literature, reading comprehension, civic education, English as a foreign language and French as a foreign language (table 1).

The main aim of the project is to examine the productivity of the various educational systems of participating countries relative to input factors--to evaluate what different school systems and their teaching methods and curriculum contribute to student learning. Thus, the primary objective was not to compare educational achievement of the countries but to relate input factors (such as the social background of students, teacher competence, curriculum characteristics and teaching practices) to outcomes (in terms of student achievement and attitudes).

At the same time, the study is intended to facilitate the comparison of different education systems. Comparisons have previously been descriptive analyses of systems and methods of education within nations. The IEA study has attempted to go beyond that procedure and add empirical

Table I  
1  
COUNTRIES PARTICIPATING IN IEA PROJECT

	Science	Literature	Reading Comp.	Civic Ed	English as a For. Lang.	French as a For. Lang.
Australia <sup>2</sup>	X					
Belgium (Fl) <sup>2</sup>	X	X	X			
Belgium (Fr)	X	X	X		X	
Chile	X	X	X		X	X
England	X	X	X			X
Federal Republic of Germany	X			X	X	
Finland	X	X	X	X	X	
France	X					
Hungary	X		X		X	
India	X		X			
Iran	X	X	X	X		
Ireland				X		
Israel	X		X	X	X	
Italy	X	X	X		X	
Japan	X					
Netherlands	X		X	X	X	X
New Zealand	X	X	X	X		X
Scotland	X		X			X
Sweden	X	X	X	X	X	X
Thailand	X				X	
United States	X	X	X	X		X

1. Poland and Rumania supplied information about science teaching in their countries. Poland tested in science but data were received late to be included in the analysis.

2. The Flemish-speaking and French-speaking areas of Belgium are treated as two separate countries because of their dissimilarities.

evidence of the outcomes or "products" of the systems.

According to Husen, the researchers who originated the project hoped to look at different international educational practices and at the variation in outcomes of educational systems, and then to identify factors which caused those differences.<sup>1/</sup> The analysis would then involve an examination of factors associated with differences between countries in the achievement levels of representative groups of their students. The feeling is that international data are valuable in broadening understanding of relationships between variations in educational practices and subsequent educational achievements. Practices to be compared are not as well represented within a single country as within a number of countries. Further, it was felt that it may be desirable to test whether a relationship found to exist in one country is a more universal one. In addition to studying productivity of educational systems, the IEA sought to develop measurement instruments that could be used for international comparisons and to establish international standards in basic cognitive areas at different levels of schooling.

In essence, then, the IEA studies are inquiries into national systems of education and are based on the premise that some factors of a society and an educational system can be used to predict and/or explain particular features of a nation's pattern of educational achievement. In the effort to increase understanding of the education phenomena, three-pronged, cross-national comparisons were made: comparisons of levels of performance of students within countries; of differences between schools within countries; and of differences among countries.

---

<sup>1/</sup> Vol. I, Foreword.

This chapter will briefly review general procedures and findings of the IEA study. But its primary focus will be on the implications of the project for the developing countries and on the methods, techniques and results of measuring educational outcomes that may be applicable to those countries.

### Organization and Procedures of the IEA Study

About 15 years ago, IEA began the cooperative effort to develop internationally valid test instruments to measure achievement. One of the first efforts was a pilot project which sampled educational achievements of 13-year olds in 12 countries between 1959-1961 to test the feasibility of cross-national comparisons. Next was an evaluation of outputs of mathematics teaching in 12 countries, completed by 1966. Two populations were sampled: 13-year olds and pre-university students.<sup>2/</sup>

The current six subject survey is a vast and complicated undertaking involving some 300 experts, about 258,000 students, 50,000 teachers, and 9,700 schools. Fourteen different languages were used in instrument construction and testing.

Three age groups were tested:

Population I	10 year-olds
Population II	14 year-olds
Population IV	students in the last year of secondary school
Population IV's	for science only: students attending school specializing in science

---

<sup>2/</sup> The results of that study are described in Husen, T., International Study of Achievement in Mathematics.

Among the instruments constructed and used were achievement tests for each subject; accompanying questionnaires for students, teachers, school, and the national centers; and descriptive and attitude scales.

Guidelines were drawn up for selecting schools and groups to be sampled. Then procedures were developed to assure that the samples drawn would be representative and that standard errors could be estimated. In an effort to give some longitudinal flavor to the tests, anchor items (common to tests for two or more populations) were used to permit comparisons between countries on growth from one population to the next.

#### The Six Subjects Studied

By May 1973 analyses had been published for three of the six subject areas: science, literature, and reading comprehension.

Science--19 countries participated; this includes all countries in the IEA project except Israel. Science was defined generally; each test included items from the earth sciences, biology, physics and chemistry. Tests were based as much as possible on common elements in the school curriculum of the participating countries, and final test items were determined by international item analysis derived from pretesting the cells in the subject area/behavioral objective grids, the availability of pencil and paper practical items, and need for anchor items between populations.

Literature--Literature tests dealt with both achievement in literature and response to literature by students in Populations II and IV. The major reason for including this subject stems from the fact that literature has a central position in the cultural life of a community. Ten countries were included in the survey. It was felt that cross-national

evaluations of achievement in literature are possible when dealing with ability to read texts and with patterns of response--that is, with non-national aspects. The IEA research sought to shed light on the relationships among facets of achievement and to better understand the influence of student characteristics, student background, curricula, and instruction on that achievement. There are broad differences between nations in the way literature is approached and taught. There appear to be no universal standards of achievement; instead, standards are relative to the aesthetic and cultural presuppositions of the nation.

Reading Comprehension--Reading comprehension achievement was surveyed in 15 countries. An earlier feasibility study showed that the international variance when compared to the intranational variance was smaller for reading than for other test material. There was generally more similarity of the objectives of teaching reading across countries than with the other subjects. However, the preparation of genuinely equivalent tests in the subject was the hardest. The study focused on cognitive aspects of reading and did not attempt to assess aesthetic or affective aspects.

Still in preparation are the analyses of the other three studies: civic education, English as a foreign language, and French as a foreign language.

Civic Education--The study of civic education attempts to measure cross-national effects of formal courses dealing with political subjects and to look at individual processes of feeling, thinking, and behaving politically. Important common cross-national curriculum elements were identified; then cognitive and affective measures were developed. Nine countries participated.

English as a Foreign Language--Eight countries participated in this study involving Populations II and IV. Common areas of teaching practice and national aims were identified at the outset. Content of the tests was influenced by different cultural settings or orientation as compared to those of English speaking peoples. But, peculiar to foreign language studies, the international test could not take into account linguistic differences between native languages of each country--these differences are likely to be one of the most important variables affecting achievement.

French as a Foreign Language--This is a second study of achievement in foreign languages. Since English has a somewhat unique position as being a principal second language or the most important foreign language in most countries participating in the IEA project, it was felt that study of outcomes of teaching of another foreign language could provide valuable insights and comparisons. Objectives of teaching French appeared to be similar across countries; consequently, a fairly standard battery of tests seemed appropriate. Seven countries participated.

### Tests and Test Construction

We have already noted that one of the primary concerns of the IEA study was the development of measurement instruments. For some time there had been a need for tests that could be used to evaluate technical assistance programs in education in the developing countries--evaluation of student competence in subject areas instead of head counts of students and graduates.

IEA spent several years developing and constructing tests for each of the six subject areas. This process had four stages:

- (1) construction of test instruments;
- (2) pretesting and revision of instruments;
- (3) dry run testing; and
- (4) final testing (and analysis of results).

In constructing its tests, IEA took precautions against undue cultural bias. For each subject area there was an international committee composed of subject matter specialists, teachers, test developers and curriculum specialists. Each participating country had a National Research Center and subject committees. The international committee for each subject was responsible for construction of the tests for its field; and the individual countries analyzed curricula, proposed item material, and conducted trial runs of test items.

### Science Testing

Cognitive Tests--Items for the science tests were selected and grouped in terms of the content and behavioral objectives they measured: functional information, understanding, application, higher processes (including analysis, synthesis and evaluation).

Phase one of science testing consisted of draft test construction and pretesting. The International Science Committee sent rough drafts of items to national centers for comment. The tests were then revised and sent back to centers for pretesting along with a manual of pretesting procedures and instructions dealing with possible sources of difficulty, such as translation, use of popular and scientific terms and substitution of local plants, animals and materials for unfamiliar ones in draft test items. National centers could substitute local alternatives as

long as they stayed within the general framework of the item and if they informed the International Science Committee and submitted translated copies of their revised items.

The next step was preparation by the National Centers of an analysis of items and tests as a whole based on results from pretesting random samples of populations.

The final tests were then compiled on the basis of (a) these analyses, (b) subject area and behavioral objectives, (c) availability of pencil and paper items, and (d) need for anchor items. Again, National Centers were allowed to change translation and presentation of items where deemed necessary as long as the International Science Committee was informed.

Attitude and Descriptive Scales --Through attitude and descriptive scales, the International Science Committee sought information on the influence of two recent developments in science teaching: first-hand experience and original investigation. Attitude and descriptive instruments had to be produced for countries with varied social and educational traditions as well as different degrees of progress in economic and technological development.

Descriptive scales were used to obtain information from the students about such school variables as methods of learning science (from textbooks or practical experience) and laboratory learning (whether structured or unstructured). Attitude scales sought information about interest in science and attitude toward school science and toward science in the world.

Development of instruments was similar to that for cognitive measures: from a large number of items, elimination and revisions were made after preliminary testing in three countries and then further pretesting in other participating countries.

Both the descriptive and attitude scales were included in a Science Questionnaire completed by students taking science courses.

It is interesting to note that an optional test of practical abilities which used simple apparatus was given to a sample of students in two countries. From data collected, it appears that these tests measured skills and abilities considerably different from those measured by tests without apparatus.

### Literature Testing

Measurement techniques were developed to measure the ability to read literary texts and to answer comprehensive and interpretive questions. These two measures of comprehension and interpretation were the cognitive measures. In addition, a questionnaire dealt with interest in literature and degree of transfer between what a student reads and his/her life-- these were the affective achievement measures.

The analysis compared achievement scores between nations and correlated them with the availability of printed matter and with radio and TV usage and with other behavioral characteristics.

A major part of the literature survey was descriptive--a study of student response to literary selections.<sup>3/</sup> New methodologies were developed to identify basic elements of response; test instruments were then developed as categorizing devices were empirically tested. The

<sup>3/</sup> Response is defined as the interaction between the individual and the literary work, an interaction that may continue long after the individual has finished reading. The response consists of associations, feelings, and reflections that occur as one reads a literary selection. The response is usually not explicit but indices of the response can be obtained. Volume II, p. 36.

approach was to relate student response to literary work and at the same time to relate student response both to socioeconomic patterns of the country and to the way literature is taught in schools.

In his foreword to the volume, Husen writes that a part of the pioneering nature of the study is, apparently, to leave open many questions regarding both the methodology and interpretation. The area of response is limited; the methodological approach had to be tentative; and the findings are in certain respects exploratory and non-conclusive.

#### Reading Comprehension Testing

There were three components of the reading test: reading comprehension, reading speed, and word knowledge. While it was easier to find consensus on objectives of instruction for test instrument development in reading than other subject areas, language and translation problems and text selection were more difficult.

Reading Comprehension--For the reading comprehension test, IEA decided that skills to be tested would be cognitive ones and not appraisal of style, feeling or literary technique. Test items consisted of reading passages. Selection of those passages was based on (a) suitable range of difficulty for populations tested, (b) number and diversity of items, and (c) variety of content and treatment. An overall criterion was that passages be suitable to all participating countries and not peculiar to any one culture or country. While it was agreed that selected test items discriminated satisfactorily between good and poor readers and was particularly effective for Populations II and IV, the test tasks were found to be too difficult for the developing countries that participated: Chile, India, and Iran.

Reading Speed--A number of passages of a low level of difficulty were selected for the reading speed tests in order that the test would measure fluency in mechanics of reading. Trial runs in participating countries helped eliminate ambiguous or too-difficult passages. The final test consisted of a four-minute test of 40 items.

Word Knowledge--The word knowledge test was based on word pairs, either synonyms or antonyms. Each country submitted pairs, ranging from easy to difficult, in their English translation. Then from some 300 pairs the National Committees indicated which pairs were impossible or difficult to translate into pairs of corresponding difficulty in the language of the country. Items retained were given a trial test and an attempt was made to get a subset of items that were nearly the same in difficulty across all languages. But the "final result is one in which the equivalences from language to language is suspect" (p. 33, Vol. III). Generally the range of difficulty was greater in English. Also, items tended to get easier in translation.

Percentages correct in the final test tended to be low in the developing countries. Overall the variation from country to country was greater than for the reading comprehension test. The IEA reported that test items showed good and relatively consistent discrimination within countries but were not as reliable for cross-country (or cross-language) comparison.

### Questionnaires

Another essential part of the IEA effort was the acquisition of background data. As part of that effort, IEA developed and used the

National Case Study Questionnaire and separate questionnaires for students, teachers and school administrators. These questionnaires provided information necessary to construct profiles of the participating countries. Passow, Noah and Eckstein (in press) have, in their report on the National Case Study Questionnaire, drawn up "national profiles" for the 19 countries which participated in the first stage of the Six-Subject Survey. The size of the per capita GNP varies from about 1,400 to 4,300 U.S. dollars in the industrialized countries, whereas it varies from 90 to 270 dollars in the LDC's which are in the study. The size of the non-primary sector as a percentage of the GNP is in most cases 90 to 95 percent in the richer countries as compared to 50 to 75 percent in the LDC's. The difference is even more marked if we measure the size in terms of number of people employed in the primary and non-primary sectors respectively.

In addition, data on school systems, teacher characteristics, family, and societal variables were obtained and are now being related to achievement data in a search for patterns of similarity and differences among indicators and patterns of student achievements.

The background data are also being used in an attempt to explain differences within nations (urban-rural, large school-small school, single sex-coeducational, high-low attainment schools) and to see if these differences of national educational performance are consistent across nations.

The National Case Study Questionnaire gathered pertinent information on:

---structure of school system, types of schools;

- processes of curricula change;
- pupil selection process;
- teacher training practices;
- instructional materials, patterns and nature of instruction;
- social and economic conditions (unemployment rates, elitist school, familial);
- urbanization and modernization;
- language and culture; and
- National Centers' judgments of education system.

The school questionnaires sought information on:

- number of schools participating;
- number of students per school;
- student/teacher ratios;
- area served by the school (urban or rural);
- staff (numbers of part-time, sex, age); and
- nature of courses.

Teachers were asked about:

- postsecondary schooling;
- age;
- hours of preparation per week;
- membership in professional associations;
- in-service training; and
- refresher courses.

Information sought from students included:

- family background: parents education, father's occupation, number of children in family; and

---out-of-school habits of an educational nature: hours spent reading for pleasure, watching TV, on homework.

The data were helpful in constructing a School Handicap Score (SHS)--a device used to adjust for socioeconomic gaps between countries and for the different communities from which the school drew its students. The following variables were selected to form the SHS:

(1) father's occupation, (2) father's education, (3) mother's education, (4) use of dictionary at home, (5) number of books at home, and (6) family size. Thus, regardless of the quality of the formal educational system in the LDC's, the IEA study was able, on the basis of the impact of the family background factors, to predict a large difference in mean achievement between LDC's and the more industrialized countries. Parents in the former type of countries are usually illiterate and no reading material is available at home. On the whole, the verbal environment in which the children grow up is almost entirely oral and there are rather few occasions in which reading skills picked up at school can be reinforced by experiences at home.

Thus, difference in achievement levels between students in developed and less developed countries could be expected, considering the overall socioeconomic setting for the school systems in the two categories of countries. The outcomes of the multivariate analyses tell us that the total effect of home background variables in both Science and Reading is greater than the total effect of all the school variables. Among the 10-year-olds, 35 percent of the variation between students can be attributed to family background and 22 percent to school factors, including, of course, all the instructional factors. The corresponding figure for

the 14-year-olds are 42 and 26 percent respectively.

### Assessment Results and Problems

Generally, IEA findings reinforce various national assessments of educational outcomes: that schools have less to do with achievement than the home and family background of the student. The study also substantiated other findings that boys do better than girls in math and science tests and girls do better than boys in reading and literature.

"Opportunity to learn" (in the science study) is the school variable most clearly established as being related to achievement. Other school system or school-connected variables that seem to have some relationship to higher achievement levels include:

- teacher training (science);
- controlled practical work rather than informal investigation (younger populations, science);
- students who do a lot of homework (literature); and
- students in classes that use a textbook (literature)

High retentivity (as opposed to selectivity) has little or no effect on the achievement of high level students.

The comment by the authors of the analysis on the science tests--that the IEA study has been more fruitful in sociological than in pedagogical findings--holds generally for the other two completed analyses as well. Little information and scant evidence have been provided for evaluating the contributions of different school systems, teaching methods and curricula to learning achievement.

The reading comprehension test results provided "few clues" as to what it is about the school environment that results in better reading achievement of its pupils.<sup>4/</sup> The science test results provide few guidelines for evaluating the relative efficiency of different teaching conditions and methods; indeed, the greater part of the schools' contribution toward science achievement appears to be the result of decisions about providing opportunities to learn and how much effort (homework, years of study) is demanded of the student.

The limited results of the study point not only to the importance of factors outside the school in determining the achievement of students in a given school, but also to the need for greater research efforts.<sup>5/</sup> Limitations on methods and data are apparent. Analytical procedures may be inadequate with the result that the effects of the home background on learning are overstated.<sup>6/</sup> Furthermore, without a longitudinal study of student learning, it is difficult to assess school influences.

A number of unsolved problems appear to have influenced test results. Science test questions, for example, were generally difficult for students in most countries. This may have been caused by constructing tests to fit a generalized curriculum rather than the curriculum of any one country.

---

<sup>4/</sup> Thorndike Speech and Vol. 2, p. 9.

<sup>5/</sup> Ibid.

<sup>6/</sup> Platt.

Translation problems were not completely solved in the reading comprehension test--there was so much variability in the data as a whole that questions are raised about the comparability of tests translated into 14 languages. There was less consistency in test scores at higher age levels than lower due in part perhaps to the more difficult ideas and more complex language of the texts. And translators may not have fully comprehended the ideas of the original text.

Cultural and national differences cannot be completely accommodated in cross-national tests. In the reading tests, the analysis points out that cultural differences appeared in interpreting motivations and emotions and that climate and style of life could have been basically different from the English pattern from which most of the test passages came.

Ambiguities and inconsistencies that appear in the study lead to questioning of the adequacy of test instruments, whether the achievements being measured are the real objectives of education, and ways to measure contributions of non-formal and informal education. These and many other unanswered questions brought to light by the study may lead to new national and international assessment efforts.

Assessment Results and the LDC's--While, as noted above, the developing countries will find useful the IEA experiences in preparation of measurement instruments for countries with a wide variety of cultures, different stages of development and disparate school systems and educational objectives, unique problems arose concerning the four participating LDC's.

Comber and Keeves in the science analysis stated that the level of economic development and lack of a long period of universal education showed up clearly in the test results.<sup>7/</sup> So great were the differences that cross-country test results of the four countries were compared, separately, with themselves, not with the 15 developed countries. A separate mean score was calculated for the four.

Results from Reading Comprehension tests in the LDC's seem to indicate a relatively greater importance of school factors compared to home influences in the developing countries than in more developed ones.

The developing countries had significantly higher student/teacher ratios and shared common problems of rapidly expanding school systems and instructional methods. And while the IEA studies point to the need for additional work on developing instruments to measure educational outcomes for non-western societies and developing countries, they also highlight the very great problems involved in developing measurements of educational achievement in general and international measures of achievements in particular.

#### The Tests and Their Validity for the LDC's

Chile, India, Iran and Thailand participated in the Science survey and all but Thailand participated in Reading Comprehension. Chile took part in Literature, English and French studies and Iran in Civics.

Up to now no representative comparative information with regard to student competence in the developing countries has been available.

---

<sup>7/</sup> Vol. I.

The IEA study is the first attempt at qualitative comparisons between developed and developing countries made according to an agreed-on international yardstick. Furthermore, the reading comprehension survey is the first comparative assessment of literacy levels among representative groups of students in LDC's. The non-industrialized or developing countries are consistently far behind the industrialized in average achievement over subject areas and levels of schooling. In Science the LDC's score was roughly one standard deviation or more below the more developed. This means, then, that in Science, the average student in a LDC scores between the 10th and 12th percentile in a developed country. The difference is even more pronounced in Reading Comprehension, where only 5 to 10 percent of the students in the LDC's score at the level of the average student in a more developed country. Chile participated, as mentioned above, in the survey of French and English as foreign languages and Iran in Civics. The mean cognitive scores in both cases turned out to be on the same relative level as in Science and Reading.

What explanations can be advanced for such big differences? First, we must emphatically caution against any premature conclusions about the "productivity" or "efficacy" of the school systems in the two types of countries on the basis of the mean scores presented in table 2. The differences that we find between the industrialized countries are negligible in comparison with the gap between the two categories of countries. There is, however, no reason to believe that the more developed countries are all on the same level of "efficacy" with regard to their school systems.

A plausible explanation is that the tests are not doing justice to the children in the LDC's. The tests might draw upon knowledge and learning experiences that are more predominant in the rich countries. Furthermore, the test situation and the format of assessing the outcomes of learning might imply a certain cultural bias against students in LDC's. A number of IEA researchers feel that while they cannot entirely refute such hypotheses, the empirical evidence does not give much support to that notion. In the first place, the content of the tests, i.e., the individual test items, went through a long procedure of scrutiny and try-out before they were "passed" by all the national subject area committees and included in the international tests. Secondly, the rank order of difficulties of items tended to be highly correlated over countries, which indicates that differences in total scores between countries are not so much accounted for by differences in particular sub-areas or topics of a particular subject as by systematic differences in level of competence. The teachers were asked to rate, on a four-point scale, each item in the tests with regard to what opportunity the students in his or her class had had to learn the subject matter that was assessed by the item. As far as Science is concerned, the average opportunity tended to be somewhat lower for Populations II and IV in the LDC's (see, Comber and Keeves, 1973). But these differences in opportunity can by no means explain more than a small portion of the difference in mean performance.

Other researchers and analyses are not so sure that there is no bias. Volume III,<sup>7/</sup> for example, indicates that some test items used

---

<sup>7/</sup> p. 148.

during trial run were shown to be too difficult for students in developing countries but were still retained. Platt feels that much research is needed to develop instruments for measuring educational outcomes in ways appropriate to the needs of non-western and developing countries. IEA's achievement tests were not a satisfactory fit for the students in the LDC's participating.<sup>9/</sup> Platt (and others) point out a number of problems:

- the tests were developed and edited in English and then translated into the language of participating countries;
- tests may have depended too much on reading ability and, based on the high rates of error on the reading comprehension tests in Chile, India and Iran, students in those developing countries may not have been able to read items in the Science tests;<sup>10/</sup>
- the concepts and values sampled in the tests were not oriented to non-western countries;
- North American and West European education orientation dominated test construction; and
- developing countries have had less experience with the multiple-choice format used.

Further research on the data and additional interpretation will undoubtedly be forthcoming.

---

<sup>9/</sup> Platt (Paper for Harvard meeting).

<sup>10/</sup> Husen.

It would be useful when evaluating educational outcomes in an international context, particularly when non-Western nations are involved, to look more deeply into the role of social and political patterns of behavior. Then achievement or competence in various subject areas could be interpreted in relation to social environment (including sex and class distinctions and family structure); to political structure and attitudes; and to education-related services (school lunches, health care and medical services, etc.).

#### Other Products of the Study

In the final analysis, among the most useful outputs of the IEA study are the processes and procedures developed and the data acquired. The international cooperative machinery and processes provide a base for continued and improved international assessments. And new research on, and re-evaluation of, the data may prove fruitful.

For all countries and the developing countries in particular, a number of strategies and techniques used in the IEA survey could be utilized routinely. These include:

1. methods of analyzing national curricula in terms of the goals to be achieved;
2. techniques for constructing instruments to measure progress toward achieving those goals;
3. procedures for drawing probability samples from target populations under consideration;
4. experiences in data processing that are appropriate for nationwide evaluation surveys;

5. routines for data collection in the schools. This includes experience in developing, distributing, and processing questionnaires;
  6. manuals developed as part of the preliminary work can be used as guides for countries seeking to improve their testing. IEA manuals provided instructions for national and local test coordinators and for individuals giving the tests; and
  7. data has been acquired, much of it yet to be analyzed.
- As part of the mathematics study, a data bank was developed. Described as a technical breakthrough, it permitted data to be packed on a single tape. With the larger six-subject survey also to be packed in an equivalently small space, it will be possible to locate the bank at several places in the world. Furthermore, when other data are added to the IEA data bank, researchers will be invited to evaluate methods and approaches to teaching and to reanalyze and reinterpret the data.

### Conclusion

One might well raise the question of the worth of an elaborate exercise like the one pursued by IEA to develop international standards of evaluation, considering the tremendous differences between the two categories of countries in terms of culture and tradition. But if the goal in the LDC's is to achieve "modernization," i.e., among other things, to create an infrastructure of knowledge and skills conducive

to an economic development which has led to affluence in the industrialized countries, then there is much to be said for attempts to measure, for example, basic reading skills and the knowledge in science that is fundamental to the creation of a modern technology.

## CHAPTER XVI

### CORRELATES OF EDUCATIONAL OUTCOMES:

#### SOCIOECONOMIC STATUS AND OUTCOME MEASUREMENT\*

##### I. Outcomes of Education and Their Correlates in Highly Developed Countries.

One might imagine an ordinary literate person finding himself curious about the easily noticeable differences among children's performance in what might loosely be termed their education, and wondering what we really know about the origins of these differences. Given the time and the purpose, he could begin to review whatever literature was conveniently available. After a short time, we suggest that four conclusions would take form; the more thorough the reading, the more

---

\* The material in this chapter was prepared by Ezra Glaser and draws, in part, from the report, "Associations between Educational Outcomes and Background Variables: A Review of Selected Literature," by Edward C. Bryant, Ezra Glaser, and Morris H. Hansen of Westat, Inc., prepared for the National Assessment of Educational Progress, under the support of the U.S. Office of Education, November 1973. Other studies include Harvey A. Averch et al., "How Effective is Schooling?" Rand Corp., Santa Monica, Calif., March 1972, pp. 1-222, and Frederick Mosteller and Daniel F. Moynihan, editors, On Equality of Educational Opportunity (New York: Vintage Publishers, 1972).

deeply entrenched they would become.

- Numerous and varied studies have addressed the general problem.
- Not too much of the amount of individual differences can be accounted for, even after diligent search (referring to "background factors" rather than classroom technology).
- There is a general consensus, nevertheless, on the "background" factors that affect educational performance.
- In a more technical context, there are very many choices about how to proceed with studies of this type; but the results seem to be relatively unresponsive to important types of study design.

A. Numerous and varied studies have addressed the general problem of accounting for differences in the educational performance of individuals. We concentrate on those studies that seek explanations in the "background factors" of the student, the school, and the community, rather than the books, teaching methods, special equipment, and other teaching technology, except as they may be associated with the "background factors." The "background factors" are taken here to include only those characteristics of children, families, schools, teachers, and the community that are substantially beyond the control of the educational establishment over periods of relatively few years: the age, sex, and ethnic origins of the child, the education and occupations of his parents, family income, the "cultural" atmosphere of the home as represented by books, periodicals, sporting equipment, musical instruments, etc. Studies which concentrate on these factors might indeed include a good deal of influence of the characteristics of the teachers, the school, and related matters (including the cost per pupil) because of the association between the educational system and the characteristics of the community. A poor rural population will score low on the various

scalings for socioeconomic status (SES): low parental educational attainment, low-wage occupations, low family income, and a general absence of books, along with large families. In such a community, the schools, teachers, and classroom procedures are most likely to lack what the educational community regards as the advantages of the educational system in more favored communities. To the extent of this association, then, the background factors also represent important characteristics of the educational establishment and its methods.

B. Not too much of the amount of individual differences among students is accounted for by attempting to associate test scores and other evidences of educational successes with background factors. This remains true after repeated attempts to explain more of the total variability by adding more factors to the analysis in an attempt better to represent the different aspects of socioeconomic status or better to capture the cumulative effects on the student of his educational experiences (not only those in the school) up to the time of the testing. The reasons behind this observation will be set forth later.

C. There is a general consensus among the majority of the studies, nevertheless. The same kinds of background factors emerge again and again to account for whatever portion of the variability among students (and among schools) can be explained by associative models. This should not be interpreted as holding that all studies are equally valid, that loose and inexact research methodology will produce worthwhile results, or that particular studies cannot be found that contain superior results (more firmly established, more detailed, more success in isolating the effects of similar influences). But the general tenor of the results

are sufficiently similar to impress the nontechnical reader with the understanding that he is discovering the same sort of thing over and over. (Some of the more important differences rest upon technical points that will be taken up later.)

D. There is a bewildering variety of ways in which to approach the measurement of SES, which all seem to insist be a part of the associative model. If it is not practical to measure family income directly, which of the many surrogates to use? If he takes the occupation of either or both parents to represent income, which classification system to use and whose scaling? How to represent the cultural climate and intellectual stimulation of the household? Should any specialized cultural and artistic interests be sought, by asking about musical instruments, artist materials, or athletic equipment? Beyond the choice of variables, what populations are to be probed? How to sample? How large a sample? What procedures for the gathering of data? What kinds of validations and quality controls? All of these questions represent the distinction between a review of research results and the design and conduct of a study.

With these introductory comments out of the way, we turn to some methodological matters that are essential to the understanding of associational analyses and the models that utilize them.

The basic idea of statistical analysis is to enhance understanding about variability. If all students performed alike, there would be nothing to study. We start, therefore, with the notion of total variability--the differences of all of the individual scores from the average of the total population. (For reasons not explained here,

the measure used is the sum of the squares of the differences.) The basic approach is to partition this measure of total variability into components that assist in the understanding of the "origins" of the variation. (The word is used loosely here; all that the model postulates is association, rather than causation. If causation is ascribed to a statistical relationship, it rests upon considerations other than the decomposition of total variation into useful fragments.)

We assume that a large sample of 12th grade students has been tested so as to eliminate concern about sampling error. Suppose the averages are as follows (the data are hypothetical):

	Race		
	White	All other	All races
Sex			
Male	64	56	63
Female	60	52	59
Both sexes	62	54	61

There are obvious differences in the average performance of the sexes and the two categories of race because they vary from the overall average score of 61.

If a particular male white student achieves a score of 70, one can identify 70-61, or 9 score points, as the amount of his total variation from the mean of 61. Of that total variation of 9 points, three points (64-61) are "explained" by the fact that he is a white male, and the remainder (6 points) is unexplained by the sex-race classification.

The above example is an illustration of a means of explaining some of the sources of variation in an individual student's score. A way is needed to summarize such variation over all students. Because of properties that need not be discussed here, sums of squares of difference are used. It can be shown that:

- A. The sum of the squared differences between individual-student scores and the overall mean is equal to:
- B. The sum of the squared differences between the individual scores and the individual cell (sex by race) means, plus
- C. The sum of the squared differences between the cell means and the overall mean (summed over all students).

The ratio (C/A) is the proportion of variance explained by the background variables--in the above illustration, by sex and race.

In the real world, one does not have test scores for all students in the universe, so there is sampling error in the estimation of the means. Also, even if he had scores for everyone, there would be measurement error that would tend to distort the means. Such sampling and measurement errors tend to make estimation of the proportion of explained variance less precise, but the additive relationship given above still holds.

In particular, such methods can be used to measure the amount of variation of students within schools--i.e., that cannot be attributed to the differences among schools--and also to measure the differences between schools.

There are two reasons for introducing the distinction: differences between school means and differences of the individual students from the mean scores of their own schools. First, the procedure makes a crude separation into variables dealing with the school (and, by association,

with the community and whatever the students have in common), and those variables which are distinctly descriptive of the individual student. Second, it is usually found that the proportion of variance (variation measured by the sums of squares of differences) among school means of achievement scores associated with school averages of SES variables is larger than the proportion of variation of student scores explained by SES (ignoring differences among school means). In comparing results of different studies, it is necessary to observe whether the school or the student is the unit of analysis. (Note that the average of all of the individual scores for a school is a school variable rather than an individual variable, even though the basic experimental data take the form of individual scores.)

There tend to be high intercorrelations of different variables that represent a single factor. Hence, for example, there are persistent relations between family income and a variety of characteristics of the housing, parental education, parental occupations, number of rooms per member of the family, number and vintage of automobiles, and other items. Within a particular study, these individual variables frequently correlate in similar ways with educational achievement and other educational outcomes. However, the introduction of a variable from another kind of factor (say, the sex of the student or some measure of his ability) will typically add to the proportion of the total variance explained. When there is "confounding" in the analysis because of these intercorrelations among the "independent" variables, there is no statistical procedure for separating their effects. Addition of further variables strongly related with family income will add little to the explained

variance. It would be an error in interpretation to credit the first variable (with the order chosen arbitrarily) with high explanatory value, and the others with minor explanatory value. The first variable used (whichever it is) will explain a significant portion of the total variance in scores.

Common sense would seem to insist that there are important differences in the innate ability of people, after allowance is made for any influences that might be attributed to their sex, race, age, home environment, school conditions and possible other factors. No way has been found to measure innate ability uncontaminated by acquired knowledge.

Project TALENT collected 14 measures of ability from the sample of 2,900 12th grade males. They ranged from two information tests--practically entirely acquired knowledge--through a group of intermediate items, to visualization in two dimensions--the most abstract measure of individual ability. SES measures explained .27-.28 of the variation in the information tests; it decreases through the sequence to .06 for the least knowledge-oriented tests. Even the most abstract tests contain the risk that they are influenced by both learning and the practice of taking tests. Bachman used nine tests of ability in his study of 10th grade boys; after considerable analysis, he decided to use a "Quick Test" of word-and-language skills. Perhaps a substantial part of the unexplained variation in achievement after accounting for SES and similar factors can loosely be attributed to "individual differences" consisting largely of the dimensions of ability and motivation.

Earlier school performance has sometimes been used as a measure of innate ability. True, performance in sixth grade (test scores, grades,

teacher appraisals) is a very good predictor of 7th grade performance. But the 6th grade performance already contains a mixture of ability, motivation, self esteem, and other factors that represent an accumulation of knowledge, attitudes, adaptations, and preferences that could hardly be considered innate in any useful sense. Indeed, the use of such predictors explains little except the likelihood that students tend to be consistent performers: good students (however assessed) are likely to be found to be good students next year.

We now turn to a summary of the principal conclusions concerning the influences that have been found to affect achievement scores and other educational outcomes. There are many differences in detail, and much that is in dispute both in terms of conclusions and methodology. An extensive literature deals with these differences and the dispute surrounding school vs. home in determining learning.

#### The Influence of Background Factors on Educational Achievement

Almost every major study has included some measure of socioeconomic status (SES). One can expect SES alone to account for between 10 and 25 percent of the variation in academic scores. Most common measures of SES include occupation of parents, education of parents, and items in the home.

Membership in ethnic group ("race") and sex would certainly seem to be important components of the "socio" part of socioeconomic status. Most of the studies treat race and sex of the student separately, not because they are not vital social characteristics, but because their potential importance argues that they should be separately measured.

In general, differences in achievement scores of boys and girls, men and women, are small. When specialized subjects are considered, some differences might be large enough to mention. The most important of these is the advantage boys have over girls in science scores. It seems plausible that cultural influences are at work here; it will be interesting to observe whether these differences diminish with the changing attitudes concerning occupations suitable for women.

Even after the SES factors have been statistically controlled, a significant increment of explained variance is associated with race. Race is highly correlated with measures of SES. However, the principal studies have shown that there is an additional contribution to the explanation of variance beyond that which can be attributed to SES. Studies show that race contributed an additional 5 to 10 percent to explained variation after SES and family structure and stability had been accounted for.

What is suggested here is that there are other mechanisms at work that detract from the performance of minority youth than are measured in the usual SES scalings.

A study in Great Britain reported finding strong relationships between reading comprehension scores and three parent-student variables: aspiration for the child, literacy in the home, and parental interest in school work and progress. The percentages of explained variance within schools ranged from 12 for lower junior girls to 26 for top junior boys. Other studies have pointed to similar home environment influences and to racial differences in these influences.

### School Characteristics and Educational Achievement

In the attempt to isolate the influences of school variables, we meet a familiar problem: they are correlated with other factors that are known to explain variations in educational achievement. Mayeske, using Coleman data (Equality of Educational Opportunity Study), analyzed eight student variables that are associated with the schools they attended. SES, for example, ranged from .28 to .40 of the five grades surveyed (1, 3, 6, 9, 12). That is, 40 percent of the student variability in SES scales was associated with the schools they attended. If SES has already been accounted for in the analysis, the effect of school characteristics is thereby diminished. The other variables also showed significant correlations (expressed again as squares of correlation coefficients, or explained variance): family structure and stability (.12 to .24), racial-ethnic group membership (.56 to .69), attitude toward life (.09 to .22 with grade 1 omitted), expectations for excellence (.10 to .15 with grade 1 omitted), educational plans and desires (.10 to .13 with grade 1 omitted), and study habits (.11 to .19 with grade 1 omitted). Achievement scores themselves--the variables to be explained were also highly related: 34 to 37 percent for the five grades. The warning is out: merely observing the schools will bring in a number of other statistically significant factors by the back door.

Mayeske also provides an approximate measure of the incremental contribution of school characteristics after SES has been allowed for. The combination of SES and a number of "family process" variables amounted to 48 percent of the variability in scores. The addition of

school factors brought the total up to 54 percent, an increment of six percent of the total variability in achievement scores for individual students. These results explain more of the variability than most other major studies, perhaps because of the many refinements in his analysis and the substantial size of the sample. Other large studies generally found that school factors accounted for five percent or less after adjusting for individual background factors.

A more direct view of the school influences can be offered by basing the analysis on the differences between schools. The impact that school-wide background variables can have on individual student outcome measures is limited approximately by the proportion of total variance of the outcome measures that is accounted for by the variance among school means of the outcome variable. In some illustrative cases, approximately 10 to 35 percent of the variance of the individual student achievement scores were accounted for by the variance between school means. It is this 10-35 percent portion that schools can explain, even with the inherent relations with other background variables.

#### Attitudes and Educational Achievement

Various studies have investigated the relationships between achievement scores and such affective states as attitudes, motivations, self-perception, aspirations, intentions, and expectations. Many of them have found significant relationships. For illustration, we turn again to Mayeske. Four attitudinal variables were used to attempt separate explanations of achievement scores. The variance explained (squares of simple correlation coefficients) are shown for grades 3, 6, 9, and 12: Expectations (3 to 15 percent), Attitudes toward life (1.7 to 22),

Educational plans and desires (6 to 26), and Study habits (5 to 14). In the first three, the lowest association was that for the 3rd grade and the highest was for the 9th.

It would be reasonable to expect that these attitudinal factors would be related to both SES (home and community characteristics) and the school. Indeed, Mayeske found correlations between the four variables listed above and SES, reaching to a high of .29 for educational plans and desires for the 9th grade. However, these are not so high as to suggest that most of the relationship of attitudes to achievement scores really represents an indirect way of recognizing the effect of SES.

#### Outcomes Other Than Academic Achievement

The results presented above used educational achievement as the criterion variable--the variable to be explained. It is also useful to consider other outcomes of the educational process, and to inquire into the factors that explain their variability. A simple model can be constructed which relates the education and occupation of an observed subject with the education and occupation of his father, the number of siblings in the family, and all of these factors to the income of the subject.

It is hardly surprising that educational attainment (highest grade reached) is related to occupation, income, labor force participation, unemployment (negatively related), and rates of pay. For studies made of teen-age and young adult populations, their attainment at the time of the survey is also correlated with their participation in later educational activity. It has also been established that favorable

attitudes toward school and toward work are correlated with SES as well.

The importance of these findings lies in their assurance that all of the relationships that have been investigated for the explanation of achievement scores are not isolated from the realities of later life. Indeed, the same kinds of background factors that explain success in school also promise favorable job experiences, continuation of education, and useful social attitudes.

## II. The Extension of Findings to Developing Countries.

### Background Factors and Educational Technology

The analyses presented in the previous section do not address questions relating to the improvement of the educational system. Rather they focus largely upon the kinds of students that come to the schools to partake of the educational process, methodology, and technology, whatever these might be. There is little point in suggesting that a developing country would be better off if the parents of the students were better educated, or if their home life contained more intellectual challenges and cultural opportunities, or if the income were higher.

What insights, then, come from the U.S. studies and how can they be used to view the progress of developing countries in raising the educational achievement and attainment of their people? Several answers suggest themselves:

1. The importance of differences in social and ethnic origins in the U.S. (and other economically advanced countries) warn that progress in countries with considerable socioeconomic heterogeneity

will be likely to be uneven. Those who are presently better off in SES terms will be in a position to take advantage of all advances in education with little delay and their gains will probably be substantial. Those who are at a SES disadvantage will probably not be in as fortunate a position to realize the potential gains of an improved educational system.

2. A direct consequence of the above argument is that there will be an increase in inequality of educational achievement and attainment between the more favored and less favored groups in the population. If any other result is to be obtained, there will have to be a deliberate effort (presumably designed with considerable understanding of the sociocultural characteristics of the groups to be served) with special programs and additional resources directed at the less favored groups.

3. Without truly heroic efforts, advancement of educational status of a country's population will be gradual at best. As each succeeding generation offers a better start for its children, they will have the potential to take advantage of the formal educational opportunities. Indeed, it would be easy to show that a stationary and unchanging educational system would produce better results each year as the successive waves of entering children were better equipped to succeed. Although this might generally be true, there have also been examples of very substantial gains in short periods. Two kinds come to mind:

A. The very substantial gains of the children of immigrants to the U.S. over their parents' educational status in a single generation, possibly traceable to the radical changes in social climate and expectations from the "old country" to the U.S. This involved, of

course, a shift from one culture to another.

B. Countries which initiated large scale children and adult education programs, at least up to a level of functional literacy; China is such an example: "In 1973 Chih Chun, deputy head of the Science and Education Group under the State Council, reported that there were 127 million children in primary schools and 36 million in middle schools, five or six times the number in primary schools before liberation and 22 times the number in middle schools." Also the social background of students in higher education is reported to have changed radically: "In 1958 peasants and workers are said to have comprised only 19.5 percent. At present, it is reported that more than 90 percent are of worker, peasant, or soldier origin." (Encl. Brit., 1974 Yearbook, p. 185).

4. Factors outside of the educational system can evidently prepare the country for improved educational status, notably economic factors. Improved standard of living would probably itself create new opportunities for children to learn both within the formal educational machinery and outside of it.

5. Cultural progress, not obviously related to the schools or formal education, would also be expected to influence attitudes, motivations and opportunities that would be reflected in achievement and attainment measures.

6. Health factors, concealed in the SES measures for the U.S., are well known. Surveys have repeatedly shown poor health conditions among the poorer and less educated groups, possibly reaching a climax in the migratory farm families where practically all indices of health

and nutrition point to extreme low. In many developing countries, larger segments of the population are without health services and adequate nutrition, at least as cheerless a picture as U.S. migrants. Present knowledge of the damaging effects of low nutritional standards among children suggest that real gains in education can be achieved simply by better nutrition.

None of the above suggests how the educational system might best proceed with any program for improvement. There is merely the suggestion that new target subpopulations might require special methods if they are to advance rapidly.

Considerations in Transfer of U.S.  
Results to Less Developed Countries

It is always hazardous to transfer the results of complex social studies from one country to others of greatly different character. Having offered some tentative ways of interpreting the U.S. data in the context of a developing country above, we now take up some of the considerations that might either qualify the transferred conclusions or indicate the conditions under which the transfer might be valid.

1. Social classes. Although there are great differences in the social status of various groups in the U.S., the social status itself is poorly defined. There is opportunity for upward mobility, and there are many governmental programs to encourage it, free public education being a prime example. Moreover, although the extremes are far apart, a large fraction of the population occupies a middle ground of SES, and the middle ground improves significantly with each succeeding generation. As noted above, urban children of immigrant parents have moved from

quite low SES to economic, social, and cultural success in a single generation, and in large numbers.

Many developing countries have much more rigid social classes, often with social, legal, or economic restrictions that would limit the upward progress of the less favored groups if they were to persist. Indeed, educational achievement and attainment might be more highly correlated with these rigid SES patterns than is the case in the U.S., implying a limit to educational progress unless other changes are brought about to remove traditional obstructions.

UNESCO reports of educational attainment sometimes show separate data for population groups within a country: sometimes African and non-African groups or Moslems and others. If there are large fractions of the population in groups far from the national average, one might expect little progress from the low SES groups. Children of poor, landless, illiterate, rural parents would not seem to be in a position to respond to gradual improvements in the established educational technology.

2. Statistical analysis of divided populations. In some underdeveloped countries, the population is divided into a "privileged" class and a disadvantaged class, with a void between them instead of a middle-ground population. The two groups are usually ethnically distinct. A graph of the SES of the population takes the form of a U-shaped distribution, with many people at the extremes and few in the middle. The same is true of educational attainment and achievement. Statistically, the pattern differs from that for countries (like the U.S. and other economically advanced countries) where the

population is closer to the normal distribution (heavily populated in the middle ranges, with fewer at the extremes):

The variables that represent SES are highly intercorrelated more so than in developed countries. The disadvantaged are typically poor in all respects: income, health, housing, education, possessions, occupations, life expectancy. There is almost no overlap with the distribution for the favored population, which scores high on all of these variables.

There is a high correlation between variables representing education and those representing SES, for practically any variables that are taken to represent education and SES. The absence of the middle-range values makes the distribution act like two loosely aggregated points; a line drawn through the two clusters will explain an inordinately large share of the variation of one (say, educational attainment) in terms of the other (say, SES). The result is to greatly accentuate the results of the U.S. analysis. A fairly detailed and accurate picture of the educational status of the population could be developed from their SES. (Note: The probabilistic interpretation of the correlation coefficients cannot rely on tables computed from normal distributions, or--for multivariate analyses--from normal joint-distributions.) The statistical properties are merely a reflection of what was set forth above: that special programs would be required to break through the existing patterns if the disadvantaged groups are to enjoy more advanced education.

There are probably simpler and better ways to approach the analysis of such non-overlapping subpopulations. The strongly U-shaped distributions can be portrayed by 2 x 2 tables which merely

record the average SES and the average education for each of the two groups. If the variation within each group is small compared with the variation between the groups, the simpler tabular analysis should give an adequate picture of the country without further computations.

As in the advanced countries, the relations demonstrate only associations between factors, not causality. In strongly divided countries, this nicety will not put off those planning social improvement programs. There is no need to quibble over the "cause" of poor education in the immobilized lower social classes; in some areas, schools do not even exist, and in others, they are costly.

Again, the implications for social improvement are plain. It would be unrealistic to begin with the effective abolition of social structures in underdeveloped countries with the intention of letting education take its usual course (i.e., retaining earlier relations of education and SES). Many countries have already proceeded by upgrading the educational system directly, deliberately breaking the pre-existing relationships in the face of social and technological obstacles. All that the more traditional relationships would mean is that there would be a desirable chain reaction: the higher educational status of children, compared with their parents, would lead presently to higher social and economic status. Indeed, it is probably difficult to move forward in one area without doing so in the others; whether or not the country's plans included such goals.

### III. Some International Comparisons

The comparisons of national averages of literacy rates, educational attainment, and for such SES measures as per capita income and family size are bound to be insensitive to the real relationships. Only if there were very small variations among the people within countries would the national comparisons reveal the relationships between educational variables and SES for the individuals. We shall see that the well known heterogeneity within underdeveloped countries shows up even in a summary analysis.

We present examples of great differences between urban and rural populations and also an example of the consequences of ethnic differences. In a country with U-shaped distributions of education and SES, the averages tend to fall in a middle ground where few people are found. Hence, we present only some illustrative data to show that the SES-education relationships are sufficiently persistent that they are revealed by crude analytical procedures.

The examples of countries selected below are governed largely by the availability of data from UNESCO sources, taking advantage of attempts to develop comparable data on social conditions for a number of countries. There is no intention to suggest that any country is praiseworthy or blameworthy because of its inclusion. Kenya, for example, is the only nation for which a direct comparison could be made of African and non-African populations.

The UNESCO definition of literacy: "...ability to both read and write is used as a criterion of literacy; hence all semi-illiterates--persons who can read but not write--are included with illiterates. Persons whose literacy/illiteracy is undocumented are excluded from calculations; hence

the percentage of illiteracy for a given country is based on the number of reported illiterates, divided by the total number of literates and illiterates." (UNESCO Yearbook 1970, p.31). The standard form of the statistic, when it is available, is the percentage of the population ten years of age and older who are illiterate.

A second measure of base-level education used by UNESCO is the percent of the population 25 years of age and older with less than 4 years of educational attainment. Where both of these statistics are available for a particular country, both are included in the tables in this section.

SES can be represented by per capita income (measured in U.S. dollars) and by average family size, also from UNESCO. It will be remembered that these measures are frequently used to study the relationships between educational variables and SES within the U.S.

UNESCO and the other organs and departments of the United Nations have labored long and diligently to develop comparable statistics on the economic and social characteristics of the various nations of the world. In citing UNESCO data, we merely use them for illustrative purposes; no attempt has been made to appraise their quality. It is well known that they vary in quality from one country to another, and that there are components of the population that are difficult to measure accurately in most large underdeveloped countries. For this reason, the tables presented below do not attempt to use countries where the relationship being illustrated rests upon relatively small differences in the reported data.

Table 1 contrasts some countries with relatively small portions of the population lacking 4 years of educational attainment or literacy (UNESCO definition), or both, with examples of countries at the other end of the scale.

Each of the better-educated countries reports family sizes averaging between three and four persons. Per capita income is reported only for Australia; it is high among the countries of the world.

The situation is very different at the other end of the distribution, even restricting the comparison to national averages. The percentages of the population failing to complete four years of schooling among the selected nations is in the 80's and 90's, with two of them in the high 90's. Mali has a high-90's rate of illiteracy. Except for Romania, the family sizes range about five to six persons. Except for Kenya, the per capita incomes failed to reach \$300.

Table 2 penetrates the curtain of the national average a bit by showing some examples of countries with large differences in education between the sexes. In Table 1, Algeria was reported as being a little over 80 percent illiterate. Now we can observe that this average conceals the rate of about 70 percent for men and a bit over 90 percent for women. Similarly, Nepal's average was about 91 percent; it is composed of about 83 percent for men and over 98 percent for women. Even in a country of such low literacy, a large sex difference exists. Among the more favored countries, Bulgaria was only 9.8 percent illiterate, but even here there was a large sex difference: under 5 percent for men and almost 15 percent for women. In a more middle range, Cambodia provides the most extreme example. (These sex differences in literacy are not typical; Table 2 contains the most prominent differences.)

Table 3 calls attention to the educational differences between urban and rural cultures within the same country. Data are available for countries with illiteracy rates that might be loosely called low, middle, and moderately high. (These are the only countries with this detail available in

the Yearbook.) In all three cases, the urban literacy was far better than rural literacy. As noted earlier in this chapter, the classifications "rural" and "urban" probably stand for a variety of other socioeconomic differences as well: ethnic composition, income, family size, etc., so no causal mechanism is implied.

Table 4 reports the only country with the standard educational measures available for major ethnic subdivisions of the nation. The non-African portion of Kenya is much better off--in basic educational attainment--than the Africa portion. The difference is not substantially modified if one observes the data for females.

Table 5 suggests a way of looking at illiteracy data to detect recent changes in the education of the population, even though the data are all reported for the same date. In 1962, a quarter of the population of 15 years of age and older were illiterate. But, on the same date, those who were over 15 and not yet 20 years of age had an illiteracy rate of only half of that. These recent gains made a little more difference to the women than to the men.

In this final section of the chapter, we have given some of the indications that the association between SES and educational achievement and attainment that was established in studies of the U.S. also apply to countries that range from the least developed to the most developed, in economic and social terms. There was also the suggestion that the entire country was not generally a sensitive unit for analysis; major components of the population can be separately examined with substantial increases in our understanding of the educational situation. The within-country associations of education and SES have implications that were noted earlier in this chapter and also elsewhere in this report.

**Table 1. Some Illustrative Data from Selected Countries: Illiteracy, 4 Years of Attainment, Per capita Income, and Average Family Size**

<u>Country (and year of latest education report)</u>	<u>Percent of Population 15 years and older with less than 4 years of schooling</u>	<u>Percent of Population 10 years of age and older who are illiterate</u>	<u>Average Family Income (\$ U.S. and year)</u>	<u>Average Family Size (persons and year)</u>
<u>Some relatively high-performance examples:</u>				
Australia '66	1.0		2960 '72	3.5 '66
Bulgaria '65	21.4	9.8		3.2 '65
Hungary '63		3.1		3.2 '63
Uruguay '63		9.6		3.8 '63
<u>Some relatively low-performance examples:</u>				
Algeria '66	91.8	81.2	290 '71	
El Salvador '61	83.3	51.2	290 '72	5.2 '71
Iraq '57	97.8		280 '69	6.0 '65
Kenya '62	90.1		810 '72	4.9 '62
Mali '60-61		97.8	71 '63	
Nepal '61	99.0	91.2	80 '70	
Romania '66	81.8			3.2 '66

**Table 2. Some Examples of Countries with Large Differences in Illiteracy Between the Sexes**

<u>Country (and year of latest report on education)</u>	<u>Percent of population 10 years of age and older who are illiterate</u>	
	<u>Males</u>	<u>Females</u>
Algeria '66	70.1	92.0
Bulgaria '65	4.8	14.7
Cambodia '62	30.1	87.3
Nepal '61	83.3	98.5
Paraguay '62	19.0	31.3

Table 3. Some Examples of Countries with Large Differences in Illiteracy Between Urban and Rural Populations

<u>Country (and year of latest report on education)</u>	<u>Percent of population 10 years of age and older who are illiterate</u>	
	<u>Urban</u>	<u>Rural</u>
Bulgaria '65	5.2	13.8
Ecuador '62	11.9	44.5
El Salvador '61	28.8	66.3

Table 4. Kenya: Educational Attainment of African and Non-African Subpopulations (Percent of population 15 years and older with less than 4 years of schooling, 1962)

	<u>African</u>	<u>Non-African</u>
Total Population (males and females)	92.8	37.4
Females	97.6	46.0

Table 5. Paraguay: Illiteracy of Selected Age Intervals (Percent of population 15 years of age and older who are illiterate, 1962)

	<u>15-19 years of age</u>	<u>15 years and older</u>
Both sexes	13.1	25.5
Men	12.0	19.0
Women	14.4	31.3

## CHAPTER XVII

### IMPLICATIONS AT THE MICRO-LEVEL: CURRICULUM--SUBJECT AND COURSE DESIGN, TEACHING METHODS AND INSTRUCTIONAL METHODS

Thus far we have discussed a variety of outcome measures and methods for evaluating education. This chapter asks: How can outcome assessment affect the internal aspects of education, its content and process? Specifically, we address the planners', administrators' and teachers' need to be able to understand and to utilize those outcome measures to assess and formulate curriculum or courses of study and to design teaching and other instructional methods.

Program administrators and national educational planners usually determine what courses will be offered in a given educational program or by a series of educational institutions. The specific concerns include, in addition to choices in language of instruction:

- (1) What blocks of courses should constitute a primary and secondary educational curriculum?
- (2) What is the appropriate sequence of blocks of curriculum?
- (3) How similar and interchangeable or regionally unique shall these courses of study be?
- (4) What will be the mix between academic and practical training at the primary or secondary level, in

vocational training programs, etc.?

Outcome data provide information for a better understanding of these curriculum issues. When achievement test scores suggest, for example, that students readily can learn some parts of curriculum, the question arises: Should this course of study be introduced earlier?

A concomitant of course selection and design is the development and production of instructional materials. Some materials are for generalized student or teacher use. Examples include: blackboards, pencils, paper, notebooks, tape recorders, and overhead projectors. Others are highly course specific, such as particular texts, workbooks, films, television or radio programs, and science laboratories. Some materials rely on verbal skills; others offer visual or auditory cues. Some are used to teach at the students, others are actively manipulated by the learners.

Understanding of the educational outcome achieved by different instructional methods becomes possible. Some methods in fact depend upon a detailed examination of outcomes intended and then a test of outcomes reached. For example, it requires detailed definition of each unit of knowledge that is to be acquired and TV programs also require far more specificity about course content. In pilot areas in Niger, where rural education had traditionally stopped at the 6th grade, schooling in grades 7-9 is now for the first time provided by television broadcasts supplemented by monitors, who are primary school graduates trained for three months (Tickton, 72). Outcome measures are made more difficult because media use enables the classroom to move out of the school into community centers or individual's homes. A

number of countries have media programs, usually via radio by which individuals, who live in remote areas where secondary or specialized schools do not exist, may receive instruction.

What is the way in which outcome measurement can inform the process by which the teacher and the learning environment influence the students? When outcome measures are available, the different teaching methods can be studied. A brief taxonomy of traditional methods includes the lecture, group recitation, group discussion, laboratory or practice, and project methods. Each of these methods creates a learning environment which affects the efficiency with which information is absorbed and the way students learn "how to learn." Each varies in the degree of:

- (1) student to student interaction;
- (2) student-teacher interaction and feedback; and
- (3) opportunity for self-discovery, student participation, individualized pacing, and discipline.

The present interest in a multipurpose education serving a variety of learners, in a wide range of learning environments, has stimulated experimentation in teaching methods that look to outcome data to judge method performance. Non-formal educational programs, in particular, have opened the door to new methods in such areas as learner-teacher partnership and co-determination of objectives, content and materials.

The major concerns of educators interested in methodology that can be better addressed when specific data on achievements of students become available include:

- o Which methods are most suitable for each type of learner, e.g., the slow learner, the older child, the adult, the

very young?

- o Which methods are most suited to each of the various learning environments?
- o Which methods convey certain subject matter better than others?

Formulation of curriculum requires scientific data and information on the degree and kind of learning actually occurring. The output measures mentioned in this volume can be used to distinguish between the influences of some component of curriculum. Information gathered from these measures can be used to explore:

- o If students are ready to handle the material presented; do they have the proper foundation in skills (i.e., reading) and information.
- o If material is presented in a logical orderly fashion.
- o If the methods of presentation are suitable to individual differences in ability and are consistent with regional, national or tribal cognitive styles.
- o If the method encourages memorization of facts or comprehension of principles.
- o If the subject matter is suitable to expected conditions of employment and living.
- o If the materials used, such as laboratory equipment or farm tools, facilitate learning.

Criteria are established for each type of educational program. In assessing attainment of those criteria, it is desirable and potentially possible to gain insight into the effects of the process

variables. However, before proceeding to an indepth exploration of the applicable measures, there are two considerations of importance which often make it difficult to pin-point contributing curricular strengths and weaknesses.

(1) The recent findings of many studies of educational effects, from the I.E.A. assessments to specific program evaluations, all conclude that the home background, the socioeconomic status of the student and the student's environment have enormous impact on student performance. Thus, when looking at outcomes, it is necessary to have some background data on the type of student involved. For example, when comparing students from two different schools who study a course in chemistry which differs in course organization, it is important to determine whether the variance in outcomes is explained by the in-school variable or student background difference.

(2) The effects of particular in-school variables are often difficult to isolate. (In this chapter, we treat three relative to curriculum: subject matter, method and materials; but there are others such as school facilities, program organization and administration, class size, difference in class, culture or race between teacher and students, etc..) This is especially true in the natural school setting. Even carefully controlled experimental studies and pilot projects encounter difficulties. Where single factors cannot be isolated, clusters of interdependent variables may possibly be identified.

### Standardized Testing

Of the two main types of standardized achievement tests mentioned in Chapter IX--norm-referenced and criterion-referenced--it is the

latter which is most useful for curricular assessment. The standardized tests which are normed are useful for teachers and administrators in determining comparatively how well on a scale a child, school or program is doing at some level. The criterion-referenced tests are developed on the basis of a specific formulation of educational purposes in a field of study. Test items are designed to capture the range of relevant information covered and to measure student progress toward mastery of that information. These tests can be used in several different ways.

(1) Program evaluation:

Matched groups of students are taught some course in which each group is subject to a different curricular variable--i.e., lecture vs. discussion method, T.V. program vs. live teacher only--or an emphasis on inductive vs. deductive reasoning processes. The standardized test administered at the conclusion of the course may demonstrate better overall achievement on the part of one group. It may also indicate varied comprehension and mastery. One group may outperform the other on questions stressing application of principles to new examples while the other group may show greater recall of specific information. With careful analysis, these results may indicate which of the two curricular approaches are better, and in which ways. Where comparable controlled experimental programs do not exist, the accuracy of locating the independent curricular variable may drop but close scrutiny may indicate how that variable or a cluster of them interact with certain students.

(2) Student assessment:

Student performance can be studied in light of test results. A particular pattern of wrong answers may indicate a lack of understanding of some aspect of the subject matter and may signal the instructors to make adjustments in their approach or emphasis tailoring them to specific identified gaps in knowledge. These results may signal curriculum designers that there is a weak link in course organization. The results may also demonstrate that certain students consistently outperform others when exposed to a particular method and may indicate that students from a particular environment, of a particular sex or with certain abilities and aptitudes need differing kinds of treatment. This could lead to tracking of students of differing abilities into different classes (an attempt at some form of homogenous grouping) or alert teachers to specialized student needs.

Planners, curriculum designers, and teachers can restructure courses to compensate for the test-indicated lacks or to reemphasize and utilize those approaches which give good results.

The criterion-referenced tests need not be summative. The processes of the tests can be incorporated directly into the teaching process and course design. Recent research stemming from programmed instruction has demonstrated the value of breaking down a course into modules, having a specific set of learning objectives for each unit, and using a series of tests which measure the desired objectives. Students and teachers proceed to another module or lesson only when the prior learning goals have been achieved.

Assume that there are for an arithmetic course of study at a specified grade level some 200 bits of information to be learned according to the objectives identified and test instruments tried out. Deficiencies in information about each of those 200 items may be obtained by testing. It then falls to the teacher and to the school to ask: What went wrong--the test, the attitudes and motivations of the students, the objective of the school, or what? This type of on-going testing serves as an immediate guide to student progress and to curricular effectiveness.

These tests also serve as a guide to teachers. If they are well constructed and comprehensive, they serve as an indication of the goals of course. A teacher can orient his/her teaching to meet those goals. This, of course, can have negative side effects, such as teaching directly to the test, but with proper training the teacher can learn the broad purposes represented by the proxy test questions. If teachers share in the participation of test design, there is a greater likelihood that course content and test coverage will be mutually reinforcing.

#### Teacher Made Tests and Indicators

Where standardized tests are not available or inadequate, teachers can be trained to test or measure student performance and to attempt to relate the outcomes to their own methods, the course content, timing, etc. Most teachers have an innate monitoring device which tells them when students are bored, excited, interested, comprehending, etc. These instincts are often turned off by the teacher due to frustration, a narrow perception of education or a desire to separate one's self from evaluation. These intuitions and perceptions can be capitalized

on to provide the basis for on-the-spot assessment especially if teachers are trained to look systematically for particular cues, and to interpret them.

### Work-Skill Measures

More than periodic use of pad and pencil tests may be necessary to provide a stimulus for curriculum reexamination and reform. Among other things, farming, crafts and other vocational subjects can be assessed by the extent of mastery of the subject as judged by the quantity and quality of output or by the proficiency of the process of doing. It makes no sense to evaluate a secretary's typing ability, a carpenter's lathing skill or a chemistry student's knowledge in the lab by having them write out answers to a test. Measures which can be as accurate and objective as any pad and pencil test (as indicated in Chapter VI) can be used to judge a typist at the typewriter, a carpenter using the tools or a student performing an experiment. Careful scrutiny of the process and the results of the student in such a testing situation can indicate a lot about that student's ability and about the quality of his/her training.

Although apparently most useful for lab science or the vocational setting, work skill measurement can be incorporated into the academic course as well. A course in literacy may require the writing of a short story by which knowledge of the tools and ability to employ them is demonstrated at the same time as students are given the opportunity to display imagination and creativity without test pressure.

Work-skill measures not only reflect on how well the student learned, but also attempt to grapple with a different kind of knowing--

not knowing about, but knowing how to use--and attempt to relate the two. This additional information of the processes of learning and on student performance can be fed-back to the teacher or curriculum designer as a basis for planning new courses or modifying existing ones.

### Non-Cognitive Indicators

A second question regarding the educational process, which is receiving nearly as much attention as achievement in the cognitive domain, is, "How has schooling or learning affected a student's motivation, attitudes and emotions? How has it shaped the person?" Interest in non-cognitive outcomes has increased as people recognize that schools condition individuals as well as produce trained manpower.

When school organization, curriculum and teaching materials and methods are not recognized as having impacts on students, as is frequently the case, they become part of the "hidden curriculum"-- educational content which is not planned or organized and which results in student outcomes which are not anticipated, acknowledged, or evaluated.

The school system has modernizing effects on students. A concern in some developing countries is that the changes occurring in the course of modernization and urbanization have been accompanied by alienation and psychological malaise. It is difficult if not impossible to establish a clear cause and effect relationship between these developments, but it can be said that rapid change in life style has psychological consequences.

Many observers have noted that school attendance, regardless of curriculum content, has a modernizing influence on students; some have wondered if it does not also produce a sense of alienation. Schools differ, but many if not most are characterized by large groups of

children sitting still for long periods of time, listening, repeating and memorizing. They become accustomed to quantifiable units of time, mental rather than physical activity, and verbal as opposed to practical experience. The hierarchical structure of school personnel, expectations in dress and in hygiene, division into age groups, and depersonalized relationships with teachers may be contrary to home experience.

There are several methods of measuring learner non-cognitive outcomes. These outcomes may also be looked at from the perspective of what treatment variable contributed to a particular set of attitudes, if they are administered with this objective in mind.

(a) In Chapter X, there is a full discussion of a range of scales and testing procedures which assess a person's feelings of self-control, attitudes toward one's self, and others, toward school and learning in general, and motivation to achieve. Careful application of these measures during a period of instruction can help reveal student reaction to that learning situation.

(b) Where specific tests cannot be administered, observation by a trained observer may prove effective. There have been a number of attempts to develop categorical check lists and rating scales which can aid an onlooker in his perception and judgment of a classroom situation. There are always the problems of observer bias, the disruptive effect of an outsider's presence, and variability in teacher or student behavior from one day to the next. Despite the above mentioned drawbacks, observation is a direct way to get information on whether students look happy as they play or what in fact they do when the teacher leaves the room.

(c) Direct questioning of students about their feelings or attitudes toward some particular curriculum can also be a useful measure-- as well as provide them with the opportunity to participate in the evaluation process.

Educational psychologists have found that there is no hard and fast line of demarcation that can be drawn between cognitive and affective skills. Subjects such as geography and history clearly are not simply an intellectual exercise if the version taught in school conflicts with local traditions.

Even in a relatively "pure" cognitive area, such as science, attitude plays a major role in the learning process. The teaching of science illustrates the ways in which the affective characteristics of children interact with the cognitive subject matter. The first question to be asked in planning the curriculum is, "does this material conflict with the value and belief systems of the students?" An American teacher in Nigeria once described the disbelief and dismay of college students on hearing that he had loved chemistry as a youth because the laws and reactions were so regular and predictable. The Nigerian students had always believed that the natural world was unknowable and uncontrollable (Horton). In this kind of situation, the contrary scientific or rational point of view should be presented with tact. The teacher must determine whether the attitude of the students toward logic and the scientific method is favorable, hostile, or open; whether derived from the home environment or from earlier schooling experiences, these attitudes affect future learning. The sophisticated rules of logic and other features of modern scientific method have evolved over generations.

To some degree, appreciation of scientific method in developing societies may require experiencing an accelerated recapitulation of that long history of the scientific inquiry. Patience and awareness of the need for relevance to the local society are important attributes of the science teacher. A third aspect of science instruction is whether it tends to place a higher value on "pure" versus "applied" science-- which has long-term vocational consequences for students. A fourth important point is that teachers need to be aware of the possibility of inculcating discriminatory conceptions of ability. In many cultures, girls do poorly in science and mathematics. These subjects have been defined as "masculine" activities and the girls (like most people) generally perform in line with others' expectations.

Other areas of the curriculum are designed to have direct non-cognitive outcomes, including the creative arts, civic education, physical education and health education (which encompasses mental as well as physical health). Programs outside the formal system are particularly oriented toward personal and social outcomes. Worldwide, there is a growing emphasis on preschool programs, following research citing crucial early learning and development periods for children. Many children enter the first grade already suffering from malnutrition and inadequate social, psychological and physical care. Physical neglect and malnutrition can retard the learning of a child for life; malnutrition was recently identified as a major factor in achievement by the International Assessment (IEA). Pre-primary programs are needed to supply health appraisal and treatment, particularly in the most depressed urban and rural areas. Maternal child care and health,

basic education, agricultural and vocational education programs for adults have a major attitudinal orientation--observers have noted that for a literacy or other nonformal program to have any long-term effect, the participants must be convinced of its usefulness and value (a finding which may apply also to formal education as well).

The employment of non-cognitive outcome measures can provide valuable insight into the non-academic products and processes of an educational experience. Feedback from the results of such evaluation can aid curriculum planners in designing learning situations which foster desired student behaviors and attitudes, capitalize on student interest and motivation, and can stimulate individual teacher awareness of the broadened set of educational goals.

#### Followup Indicators

The measures we have discussed up to this point are all short-term performance indicators which demonstrate interest and mastery at the point of contact with the educational program. The over-arching purposes of an education are to provide some useful way of looking at the world and some useful skills to operate in the world. Thus, no evaluation of an educational program can be complete without the feedback of outcome data that comes from data on student application of learning in subsequent activities. We will discuss three important areas where follow-up measures can be particularly relevant to curriculum planners and implementors: (1) latter-stage school performance, (2) employment, and (3) life application.

Subsequent student performance in school, culled from test scores, teacher ratings or student selection of a course of study can be part

of a curriculum planner's data base. In England, educators were curious about student college preference for the pure sciences rather than engineering. The reasons were traced back to training in the 6th Form where course emphasis was on theoretical issues (Maclure, 68).

In the United States in the 1930's, a massive study was conducted of the graduates of 30 high schools that had experimental programs. 2,108 graduates of these schools were matched with high school graduates from traditional high school programs and the progress of each group in college was followed. Data came from student interviews, questionnaires, reports from instructors, official college records and comments from others who had contact with the students. Differences between the two groups in cognitive achievement, effectiveness in solving problems of adjustment, concern for world affairs and other areas were significant, favoring the students from the experimental programs. Educators planning for these goals could have used these results as a reflection of instructional methodologies (Gage).

Planners may also want to design and evaluate curricula on the basis of what opportunity the program provides the student to re-enter the educational stream at a later date. Most formal education has a well-defined sequential pattern. Courses in one year build on an information base established in earlier grades. Students from one grade pass with the next grade or next level. Vocational and non-formal training programs are often seen as terminal. In the developing countries where non-formal programs are assuming the responsibility for a sizable percentage of some educational roles, these programs can and should be capable of being linked to the other educational systems.

If this concept of a recurrent education is chosen as a goal, it will require a number of significant changes. Planners and administrators in the non-formal and vocational programs, while not trying to replicate the formal curriculum, will have to redesign some parts of their courses so they adequately prepare students who desire it to shift into the formal system. Administrators in the formal system will have to reevaluate their entrance requirements to be able to accommodate those whose education has occurred in a different milieu.

A change in enrollment and drop-out rates, especially if they occur in a particular region or school, may signal a successful or poor combination of program elements which educators should be alerted to.

Employment patterns can be studied and employer feedback solicited. In a labor market where a range of job options is available, student job selections may reflect on the employment preferences instilled by the his/her educational training. If students will not take anything but white collar jobs, as is often the problem in LDC's, then planners can look at the types of students this attitude is prevalent among and return to the programs they participated in to see if there is not some change in curriculum possible to discourage this attitude. Also, transferability of skills can be explored. Is the educational program providing a broad enough basis for its students to find and perform work in a changing world? In Chile, followup studies were done on graduates from vocational and academic (liceo) programs. After six months on the job, the academic students were more efficient than the students specifically trained for the job. Coordination with employer needs and reexamination of the methods of teaching in the vocational programs

could lead to curricular reform (Schiefelbein).

Attempts can be made to assess direct utilization of skills or information. The best measure of literacy is the degree to which the new literate reads available material, visits libraries, writes letters or serves as village scribe. A long-term outcome measure of the success of a health care program would be changes in infant mortality, incidents of epidemics and illness in an area or utilization of available medical facilities. Where outcomes are short of targets or different from program intentions, reevaluation of the curriculum in the light of these outcomes may illuminate some of the causes.

In Ethiopia, the WOALP Rural Development program interviewed farmers after the project's completion to see if they were using techniques that had been taught. Different farmers had been involved in different programs of various duration and intensity. In this instance, good comparative data was collected on the performance of each group and then fed back to the program designers to select the most effective treatment (see Chapter IV).

References

- Gage, N.L. (ed.), Handbook of Research on Teaching, Chicago, Rand McNally and Co., 1963.
- Horton, Robin, Africa 37, 157 (1967).
- Maclure, J. Stuart, "Curriculum Innovation in Practice: Canada, England, Wales, U.S.," Third International Curriculum Conference, Oxford, Her Majesty's Stationery Office, 1968.
- Schiefebein, Ernesto, "Constraints to Change in Traditional Educational Systems: Lessons from Chile," Interchange, Vol. 2: 4, 1971.
- Tickton, Sidney G., "Instructional Technology in the Developing World," Educational Broadcasting Review, Vol. 6: 2, April 1972.
- UNESCO, Book Development in Africa: Problems and Perspectives, Reports and Papers on Mass Communication #56, Paris, 1968.

## CHAPTER XVIII

### IMPLICATIONS AT THE MICRO-LEVEL: TEACHER TRAINING AND ADMINISTRATION

#### I. Teacher Training

Teachers are the vital link between the administration and the students. It is the teacher, extension agent, instructor or trainer who is in daily or weekly contact directly with the student. They are the first to observe changes in knowledge, behavior and attitudes; thus, they have considerable information on student learning and have much to gain from a systematic and scientific way of utilizing that information to assess students and to assess their own teaching practices. Teachers may not control the program but they do often control what happens in their classroom or the learning setting. They are the interpreters of the program and the program goals. They can, in fact, affect very much whether the program objectives will be achieved. Therefore, they need to be exposed to the purposes of the program, and to become familiarized with the standards or outcome measures used to determine achievement of those purposes.

#### Existing Measures: Achievement Tests

It would be opportune during the period of teacher training, either in-service or pre-service, to introduce teachers and those studying to

be teachers to the specific purposes of the programs, in which they will participate and the criteria by which those programs are being evaluated. They can be exposed to the outcome data presently gathered to indicate progress and to methods by which the outcomes are collected. A suggested orientation for teachers might include the following steps in the process of measurement:

- why it is important to measure outcome;
- the processes of defining specific learning goals to be measured, including teacher participation in goal selection; and
- actual preparation of instruments to measure goal attainment (e.g., reliable and valid tests, good observational rating scales, interviewer questionnaires).

Most teachers at least in the formal educational system encounter some form of externally prepared standardized exam. Thus, they should learn:

- who prepares it, who uses it, for what purposes;
- how the test is given;
- how it is graded;
- how the test results are interpreted; and
- how the test has been or can be revised to reflect changes in educational objectives, new curricula or revealed test weaknesses.

This type of training can also familiarize the teacher with how to use the test. For example, teachers can ask:

1. Do I teach the information covered on the test? In other words, can and do I use the test as a guide to teaching?

2. Does the test cover the information I teach? The answer to this may lead the teachers to communicate with test designers to encourage them to incorporate content covered in class. The teachers may even be drawn into the process of test design themselves.
3. Is there a pattern of student success or failure shown by test results which I can correct? The pattern may indicate the need to change course emphasis, to spend more time with slow learners, to attempt to approach the material using a different method of teaching.

A valid and reliable test, however, is only a representative sampling of the subject matter and teaching to the test should only be one of many guides in instruction. A full exposure to the range of educational purposes of the program can assist teachers in teaching to achieve the desired results without narrowly restricting their focus to the proxy measure.

Existing Measures: System Outputs and Non-Academic Indicators

Throughout this report, we have referred to a variety of applied and follow-up measures which are most often collected by planners or administrators for systems assessment. These measures include labor force participation, subsequent schooling records, practical applications of skills on the job or in the ongoing life of the individual and/or the community. On the basis of these indicators, administrators judge the value of the program, the efficacy of the teachers, and the achievement of the students. Teachers are not always in a position to collect

the data necessary for this process. They certainly can share in the analysis of it and undoubtedly can profit from knowledge of the results.

In programs such as non-formal literacy development, health care training, agricultural improvement, etc., these follow-up measures may be the best and sole method of student and program evaluation. Teachers and instructors being trained for such programs can be encouraged to maintain contact with students to see how they are progressing in utilizing the new techniques. Where this is difficult they can be encouraged to solicit data from administrators. Incorporation of this principle of evaluation into the training program for teachers helps keep teacher focus on the actual relevant outcomes of the educational process. Where outcomes are falling short of targets, teachers may be in the best position to determine why. For example, in an adult literacy class, attendance may be good, interest apparently high and results on the class quizzes adequate, but beyond the classroom no one is using the skill. Further exploration indicates that the only available reading material is highly technical and the language of the primer is considered too childish to write a letter in. A creative and alert teacher might sense the need to change the text or write a new one, to help people write letters in class, to encourage a collective effort to secure a modest library in order to get the program to produce the desired result of functional literacy.

#### Affective Inputs:

The affective attributes and capabilities of teachers are so important to the learning process that they should be emphasized in teacher selection and training. The teacher, in formal and non-formal

educational situations, is the link between the reality of the students' daily life and the development goals of the nation. If the teacher lacks understanding of and sympathy with either, the noncognitive development of students will suffer. There are two crucial attitudinal spheres: teacher's personal values, and his or her attitudes toward students. The teacher needs to be receptive to changes, enthusiastic, innovative and adaptable to new techniques and learning materials. Also, the teacher's attitude toward the students and their cultural milieu affects their cognitive and noncognitive development. Many studies on the effects of teachers' expectations for students' achievement demonstrated that students are influenced by the opinions of their teachers. Some researchers feel that teacher attitudes are the missing element in studies such as the Coleman Report in the U.S. and the IEA assessment which seek to establish the variables associated with student achievement. An awareness by teachers of their own and their students' nonintellectual traits and needs may well in itself change the climate of the classroom to foster learning and development.

Teacher training in outcome measurement for diagnostic purposes should not neglect the area of teacher attitudes, expectations and interests and its impact on accurate student or program assessment.

#### Creation of New Measures:

Training can also guide teachers to develop their own evaluating acumen to gain insight into daily or weekly learner progress.

The move toward individualized instruction, modular course construction, and the desire for producing more effective teachers has coalesced into a training program in parts of the U.S. called competency-

based teacher education. This approach emphasizes the development of specific competencies in teachers. Rather than train an instructor to teach 4th grade in general, the focus is on specific subjects and units within subjects. Teachers are taught to define clearly specified learning objectives, i.e., the capacity to perform long division. They are then taught to define the skills and behaviors the students must demonstrate to achieve the objective. Various methods of teaching to the goal are experimented with and teachers are taught to develop tests, written or behavioral, to determine when students have mastered the process. The training consists of instructor mastery of the capacity to teach a linked sequence of learning steps.

This approach has been criticized as being narrow, yet it has the advantages of providing a highly concrete training guide and stresses the necessity for teacher-learner feedback throughout the learning process.

#### Relating Training Practices to Types of Training Programs

The purpose of teacher training, pre- or in-service, is to teach specific knowledge about some subject matter, to train teachers how to teach, and to train them how to measure the effectiveness of their teaching. Introducing the trainees to the principles and particular methods of evaluation should be relevant to the training mode. Formal pre-service or lengthy in-service sabbatical training provide the opportunity to explore the theory, practice, and history of the development of outcomes assessment. At this time, teachers may be offered the opportunity to participate in the design or reconstruction of nationally used standardized tests. They may be trained to develop a

whole program of evaluation and assessment for their region, as was the case in Indonesia, for their school, their department or program. Most in-service training is short-term (though sometimes recurrent) on-the-spot assistance. It is usually program relevant and site-specific. Trainers have the opportunity to observe the teacher in action, to note student behavior in class, and to get a feel for the community and administrative milieu. The training to use outcomes measurement should address the measures the teachers actually encounter, or those they could develop such as in-class tests which would be appropriate to the particular circumstances.

#### Pay-Offs of Incorporating Outcomes Assessment into Teacher Training

1. Teacher training can most effectively be evaluated in terms of the "good" teachers it produces. One important criterion for determining "good" teachers is the quality of the work performed, or learned by their students, i.e., student outcomes. Hopefully, teachers who are trained to use output measurement will benefit by having at their disposal a systematic basis to assess and revise their teaching practices in the light of student performance.

2. Evaluation of students and programs has often seemed an alien process, conducted by administrators or officials, who are remote from the actual learning conditions. Teacher involvement can help improve the accuracy, applicability and quality of the assessment measures.

3. Evaluation has come to be tied to accountability. Teachers often feel that it is they and not their students who are being judged and fear their job security may be endangered. Teacher participation in evaluation procedures and understanding of assessment purposes can

guide teachers toward better teaching and can provide the kind of intra-system communication which creates a supportive and reenforcing rather than suspicious working environment.

## II. Administration

To fulfill their responsibilities, administrators need an effective method of needs assessment. Because of administrators' pivotal position in the educational hierarchy, it is here that much of the outcome data can prove most useful.

A major concern of administration is the efficient allocation of the resources at their disposal between teacher and staff salaries, materials and plant to insure high productivity and low unit costs. In lesser developed countries where education consumes a large share of the national budget, the budget management is a primary, vital and often exclusive concern of the administration.

Operating within severe financial constraints and in an atmosphere of reluctance to fund without a better knowledge of what is actually being produced, administrators need a greater capacity to assess the efficiency of their programs. For this assessment, they can use the principals of systems analysis employing the investment measures mentioned in Chapter V--cost-benefit, cost-effectiveness and input-output analyses. These are quantitative measures which can be used to relate personnel and material costs to such factors as pupil attendance, enrollment, progression, and wastage.

But educators and administrators have increasingly become aware that a cataloguing of quantitative costs and benefits provides a very narrow basis for allocation and assessment. Administrators need tools

to evaluate the quality of the students to provide a more accurate input to cost accounting procedures. Thus, they require information on student academic performance and behavior. Results from national or regional external examinations can be used to compare students in similar programs, to determine how many and which types of students are progressing at an average or "normal" rate. Where external tests are not available or not applicable, administrators may have to commission university researchers and psychometricians or their own staff to design new measures.

Administrators, however, are not confronted with a fixed set of inputs to be mechanically juggled and related to a varied range of student outputs. Administrators intervene in the determination of standards and the selection of the components of an educational process. They often have to choose among differing curricula or design new curricula for their own system's needs. They select, fire, promote, and arrange the schedules of teachers. To assist them in this process and relating student performance to the influence of each of these internal components of the educational system, administrators can utilize the student output measures.

Administrators must also consider themselves and their procedures as variables in the educational milieu. Administrative procedures often pervade every aspect of the educational environment. They determine class size, length of class time and school day, disciplinary measures, location of schools, use of educational plant when school is not in progress, provision of ancillary services such as guidance, nutrition and health care, libraries, tuition and other costs to students and

many other areas. These procedures must be scrutinized to weigh their effect on efficiency of operation and achievement of better educational outcomes. For example: administrators, in one school where performance has been low, must ask themselves if curricular manipulation, better trained teachers, or more efficient tracking patterns for the students would change the outcomes or if the probable causes of below average performance lie elsewhere. It has been shown that health and nutrition are important influences on student performance, capacity, and interest. Perhaps the introduction of a school breakfast or food supplement program would in fact affect student learning more.

With regard to teacher salaries and promotion, administrators can examine their incentives and workloads in the light of teacher satisfaction and turnover and how this affects student outcomes. They may ask such question as: are better teachers promoted out of the teaching role into administration? are teachers paid according to actual teaching ability and student performance, prior educational credentials, or seniority? This is a very touchy area. Teachers' organizations have been created often with the express purpose of bargaining for salary scales protected by contracts and free from administrative sanctions and interference.

Many factors bear on the process of education and influence learning. The administrator is at the juncture in the educational hierarchy where accountability, responsibility and authority for many decisions coincide. The focus that the administration establishes, the kinds of system objectives pursued and the measures used to gauge the achievement of these objectives will not only affect the efficacy of the

management of the system, but will serve to guide teachers, parents, and students. Administrative use of tests as a measure of system outputs will tend to reenforce a test-orientation for students and teachers. Where tests are used for selection purposes, they may encourage a competitive atmosphere and foster a uniform standard of achievement. Where they are used to measure student learning, they may encourage a reconsideration of certain aspects of the curriculum in the light of individual learning needs.

On-site, in-class, academic measurement may foster an attitude that school is a self-contained system. A stress on follow-up indicators (success on the job, etc.) may promote the notion that education is merely a vehicle for accreditation. The balance between a practical utilitarian education and learning for the sake of personal development can be very much influenced by the stance of the administration as it sets objectives for the system and develops methods to measure the attainment of those objectives.

## CHAPTER XIX

### IMPROVING RESOURCE ALLOCATION THROUGH OUTCOME MEASUREMENT

The basic elements of educational and development planning were introduced and explained in Chapter II. In subsequent chapters, educational output measures have been introduced and examined in a variety of forms suggested by the distinct goals of development. The task of this chapter is to reintroduce the planning framework and suggest how outcome measures are a vital component of the planning process. This is a policy oriented chapter which is designed to make explicit the practicality of the results of previous chapters where output measures were developed. Primary emphasis will be given to systems analysis as a planning tool and, in particular, the planning-programming-budgeting system (PPBS) as a flexible planning instrument.

The need for outcome measures in a planning framework is obvious--without some idea as to what is actually produced, there is little hope of providing for future needs. This is a general, uncontroversial, statement; however, it does not give much insight into the specific way outcome measures may be used. Hence, we will try to examine specific examples of the use of outcome measures in analysis of programs.

### Basic Planning Techniques

The basic planning techniques introduced earlier are manpower requirements planning, rate of return approach, social demand approach, and the systems analysis of planning for educational objectives. Although these techniques are not specific to any particular objective of the development process, they are most closely identified with planning for economic objectives. In what follows, the use of outcome measures in each approach will be briefly discussed and a lengthy illustration will be examined.

The manpower approach is based on a static view of education for development that emphasizes the outputs of education in relation to the needs for those outputs. Surveying the education sector, the planners would try to answer the question: what are the likely changes in the numbers of persons with specific occupational skills? Using this information in conjunction with the numbers of individuals with differing skills, the expected changes due to death, retirement, etc., it is possible to estimate the manpower supply. This information on supply of manpower used in conjunction with some projections of demand allows the planner to pass judgment upon the probable educational expansion necessary to reconcile demand and supply.

The advantages and disadvantages of this approach are well known; the question arises how can outcome measurements aid this planning process. The essence of the manpower approach lies in the accurate assessment of occupational skills and the educational system's production of those skills. The courses of training include the formal school system as well as vocational training courses such as provided by vocational training centers, rural education such as Rural Artisan

Training Programs in Senegal, or agricultural projects such as CADU in Ethiopia. For accurate manpower projections, the courses of educational training should yield graduates with minimal skill levels. Outcome measures are used to assess that skill level: for example, literacy tests measure the necessary reading skills, work sample tests provide standards for manual skill, and supervisor's report may provide documentation for many non-formal education programs. In short, without outcome measures, the manpower projections cannot be accurate and may well mislead development planners.

Rate of return analysis, like manpower planning, is a planning tool aimed at economic objectives. Again, the advantages and problems with this approach have been examined previously. The essence of the approach lies in using estimates of the benefits and costs of various forms of education to form rates of return over time. The planning decisions are made on the basis of comparisons between rates of return to education versus those associated with alternative uses of government funds--generally physical capital formation.

Rates of return as an output measure purport to gauge the productivity-increasing effects of education. The problem is whether they are accurate measures or whether, due to problems of interpretation and conception, they are inaccurate and misleading. In this case, the planning tool and the output measure coincide--rates of return are guides to resource allocation and an estimate of the investment effect of education.

The social demand approach to educational planning attempts to predict the education sector requirements needed to accommodate the demand for educational services. As such, it does not directly require outcome

measures except in the very simple sense of estimating throughput of students. Indirectly it may depend on individual assessment of the results of education expressed by individual demand for education.

The techniques listed so far are rather narrowly defined planning tools. A more flexible tool is the systems approach which may use the previously mentioned techniques in a broader context. The systems approach employs carefully defined education sector or development objectives, cost-benefit or cost-effectiveness techniques, and a clear statement of the planning problem of examining and choosing among alternatives.

#### Elements of PPB Systems

PPB systems have come to be defined in somewhat different ways as the basic concepts have been adapted by governments of varying sizes, complexities of responsibilities, and concerns of staff personnel and officialdom.

We define PPB here to mean a system of inquiry about, and management of, public programs and activities by objectives. It is in essence a method of governmental programming, by objectives. The formulation and assessment of those objectives, examination of alternative programs that can achieve them, measurement of resource requirements, and accountability for program results are fundamental to the system.

We view PPB in the context of a system--a system for bringing together informative documents that as a routine process of management

can provide policy officials with more and better information. The information that would be provided includes carefully examined purposes, program costs, and potential program results in achieving specified purposes through various program options, both immediately and in the longer run. The routine of the system requires analytical documentation prior to official budget and program recommendations. And as a system, PPB requires the orderly processing of analytical work so that the timing is appropriate for the cycle of work in budget preparation.

It is a system for (1) helping to achieve management by objectives, (2) formulating programs in relation to operationally defined objectives, (3) generating new program designs and specifications, (4) assembling total program costs, and (5) analyzing those programs in accord with specified criteria for measurement.

Importantly, PPB is a system that provides an occasion for social invention and innovation. By requiring a search for options, the system opens the door to generation of options and consideration of various alternatives; at least it helps assure a more prompt and greater acceptance of assessment of new ideas.

It provides an occasion, too, for the consideration of interacting and interrelated activities that serve common purposes (both in the public and private spheres). That is, it sets the stage for interagency dialogue and communication in calling for a compilation of total costs and overall program effectiveness. For example, many government departments are concerned with educational achievement; education cannot be considered as an interest exclusive to the departments of education.

Elements of the PPB process generally are not new, but their combination and systematic application to education and other public affairs is. As indicated earlier, these elements have come to be described in various ways, but we define them in terms of structural aspects and analytical aspects, and the complementary feed-back, or accountability, aspects. In the work on these elements, a series of documents has been defined; these documents are the tools through which the PPB system is implemented. In addition to evaluation studies they are:

- (1) the program structure and statement of objectives;
- (2) program analyses (cost-effectiveness analyses) and memoranda; and
- (3) the multi-year program and financial plan.

The components of the PPB system, the documentary tools, and the processes are outlined in Illustration 1.

Preparation of the several documents of a PPB system requires:

- (1) clarifying and specifying the ultimate goals or objectives of each activity for which a government budgets money;
- (2) gathering contributing activities into comprehensive categories or programs to achieve the specified objectives;
- (3) examining as a continuous process how well each activity or program has done--its effectiveness--as a first step toward improving or even eliminating it;
- (4) Analyzing proposed improvements or new program proposals to see how effective they may be in achieving program goals;

## Illustration 1

### Components, Tools and Processes of a PPB System

<u>Component Elements</u>	<u>Documentary Tools</u>	<u>Processes Required</u>
Structural	Statement of objectives	<ul style="list-style-type: none"><li>. Formulating and defining objectives</li><li>. Formulating criteria of measurement</li></ul>
	Program structure	<ul style="list-style-type: none"><li>. Classifying programs and activities into a hierarchy</li><li>. Assigning expenditures to classification of program categories and elements</li></ul>
	Multi-year program and financial plan	<ul style="list-style-type: none"><li>. Summarizing decisions taken in output and cost terms</li><li>. Projecting program levels ahead</li><li>. Projecting workload costs of current program ahead</li></ul>
Analytical	Program analysis study (Problem definition statement)	<ul style="list-style-type: none"><li>. Defining objectives</li><li>. Defining criteria of measurement</li><li>. Formulating program options</li></ul>
	(Cost-effectiveness studies)	<ul style="list-style-type: none"><li>. Developing model for analysis</li><li>. Collecting data relevant to criteria of effectiveness</li><li>. Collecting relevant cost data</li><li>. Carrying out data analysis</li></ul>
Evaluative	Program evaluation studies	<ul style="list-style-type: none"><li>. Collecting data on program performance</li><li>. Designing experiments where indicated</li></ul>

- (5) projecting the entire costs of each proposal not only for the first year but for several subsequent years; and
- (6) formulating a plan, based in part on the analysis of program cost and effectiveness, that leads to implementation through the budget.

Program objectives--The statements of objectives of governments, and of education as a function of governments, are fundamental to program structuring, program analysis, longer range program planning, and program evaluation. Unless it is clear what outputs are being sought, there is essentially no way of knowing whether agencies are achieving them by the programs adopted and expenditures made.

Formulation and definition of purposes require that cabinet officers and ministry heads ask anew about their goals in public service. What is it that is being sought by way of results or products? Or, what needs doing and for whom? Formulation also compels an inquiry into the following areas: Why is each activity currently performed being performed? For example, what are the purposes of human resources programs? Is the main objective to raise the level of output in the nation--to increase, that is, the nation's productivity and gross national product? Or, is the objective to improve the level of living and per capita income of the poorest in the nation? Or, is the primary purpose to develop the intellectual capacity of the nation for cultural pursuits? These purposes differ, and so necessarily would the criteria by which progress could be judged. The type of program that is designed would also differ, depending upon the choice of purpose made. If all these

purposes are sought, each still needs to be identified separately and progress toward each result measured independently. Those who must decide can judge in what combination they would emphasize national economic growth, correction of the worst pockets of poverty (whether or not the national output is raised in consequence), and national cultural development, for example.

Program analysis--Central to the PPB effort is analysis of programs to assess the initial specification of objectives, to analyze the probable outputs in program results in terms of the objectives as assessed from various program options, and to measure total costs of the several options relevant to the program decision.

Planning, programming, budgeting systems, as systems, are best applied on a government-wide basis in which defined objectives apply to the entire range of governmental activity. (Government-wide is being defined here to mean an independent taxing-spending decision unit.) The government-wide effort is an attempt to gain better understanding of the range of programs and of departments concerned with the same or similar objectives. It attempts to provide a process within which ministry of education, for example, may see the scope of current services under their direction that is important to satisfying the goals of other departments or of the government as a whole. Similarly, ministry of education within this process may better comprehend the contribution that noneducational ministries make to the mission of intellectual development of a population. When economic development is a central government purpose, the comprehensiveness of approaches to structuring the PPB system's work and analysis becomes critical.

Comprehensiveness in goal setting and structuring makes program analysis the starting point of the PPB process for departments of education, especially in the developing nations of the world.

The analysis process is a unifying and comparing one. On the one hand, consequences are assessed in terms of costs, both those that are immediate and those that are implicit for subsequent periods as a result of immediate action. On the other hand, they are assessed in terms of benefits or program effectiveness. Showing costs and effectiveness side by side for various program alternatives provides new information that can make rational decisions more likely.

Analysis essentially involves a reduction of complex problems into their component segments so that each segment can be studied. Questions of fact can be subjected through this process to the test of observed experience. Those aspects of the problem that involve value judgment can be separately identified and the basis of the judgment made explicit.

On the one hand, as noted in Chapter II, a cost-effectiveness analysis may use, if applicable, many of the techniques of mathematics, operations research, economics, etc. On the other, cost-effectiveness analysis may require no more technical sophistication than the pulling together of already existing data in a meaningful and informative way. Analyses may also draw upon various technical and nontechnical studies previously done.

Recommendations made on the basis of analysis within the procedures of a PPB system are presented in policy papers termed "program memoranda." The "program memo" is a document covering one major program area or a major portion of a major program area. Its purpose is to present major

program policy findings, specific recommendations, and the reasons for these recommendations, including a summary of the analyses that have been made. It is submitted prior to detailed budget preparation.

In general, hundreds on hundreds of program problems and issues would lend themselves to detailed analysis within each department of government, and many thousands for a nation. The number of problems far exceeds the analytical staff resources even in a nation as well endowed as the United States. In the developing countries, personnel limitations are severe. It is not possible, therefore, to analyze each issue in detail at the outset, or even over several years. It becomes necessary to choose a few issues for detailed analytical study. And for some issues so selected, the period of study required may postpone its immediate use for policy decision.

#### Illustrative Example of Output Measurement in a Systems Approach

The use of educational outcome measures in a systems approach can best be illustrated by an example of a PPB solution to a typical problem. The essence of a PPBS is to bring to the fore a concise statement of the problem, alternative solutions and their costs and effectiveness, and, based on this last evidence, the possible courses of action. The contribution of output measurement may be in the statement of possible solutions or in estimating the effectiveness of proposed alternatives. As an example of the PPB planning process, we will examine the possible attacks on the problem of illiteracy. High rates of illiteracy are a common problem that has many implications for developing countries. For example, low rates of literacy may impede increases in agricultural or manufacturing productivity; or they may inhibit social or political change; or low rates may hinder the spread of family planning or health information. Within the broader context of development planning,

illiteracy can pose a barrier to the accomplishment of any of the broad range of development objectives discussed previously. In what follows, a step by step example of a PPBS will be laid out to demonstrate the technique and its application to an important problem. The example will be simple, less detailed than it would be in actual planning documents.

Step one is to state the issue confronting policymakers.

That issue may be simply stated as how best to improve literacy levels among all segments of the population. Notice that before even this first step is taken it has been determined, by whatever process is used, that illiteracy is a problem for the nation. The basic issue or issues may be stated more concretely than simply how best to improve literacy rates. It may be that there are specific programs currently under public scrutiny and the problem is which to employ.

Step two is to define the problem. In its broadest terms, the problem is illiteracy and it is measured by the illiteracy rate which may range as high as 95 percent in some less developed countries. But the problem must be examined more closely if particular solutions are to be proposed and either accepted or rejected. For example, a further question is what are the effects of the problem? In the case of illiteracy, its effects are less productivity, less social mobility, less political participation, less communication about matters relating to the well-being of the population. The cause of the problem is also important in most applications of PPBS but in this case only vague reasons may be cited--such as fundamental and deep poverty (cause and effect may be intertwined here). Another facet of interest is

determining whether the problem is uniformly spread among all segments of the population or isolated in particular social or economic groups. Obviously such a question is of major importance in proposing alternative solutions. For example, questions to be answered include the breakdown of the extent of illiteracy between rural and urban persons, among ethnic groups, between geographical areas, among religious groups, etc. In most developing countries, there is a major discrepancy between illiteracy rates for urban and rural citizens. In summary, a clear and complete statement of the problem is required before any decisions can be made.

Step three involves making explicit what was implicit before the exercise was started: identifying basic objectives. So far the objective has been the general one of improving literacy rates. Yet the question can logically be posed--improve literacy rates for what reasons? Some of these basic reasons have been mentioned previously. For example, it is often hoped that raising the literacy rate may improve agricultural productivity by aiding the spread of information and techniques. Depending upon what the basic objectives are, the policy solutions may differ and hence a clear statement of objectives is required.

Step four involves specifying the criteria to be used as measures of effectiveness in analyzing solutions. As such, this is the prior step to proposing and examining alternatives. Questions to be answered include: What criteria would capture most completely the objectives as formulated? Would administration of the program provide measurement data? What means are there for collecting data and testing programs? In short, specify the grounds upon which major alternative solutions

will be judged, making sure the criteria are complete, specific, and measurable.

At this point, output measures become important and must be explicitly introduced. Accomplishment of objectives can only be gauged by proper measurement of outcomes. And criteria are most concrete and useful when phrased in terms of specific outcome measures. For instance, if it is felt that a broad objective such as increasing the literacy rate from 40 to 60 percent is required, then the output measure and the accompanying criteria might be performance, by school leavers, on a criterion referenced standardized test. Alternatively, if the rural productivity objective of literacy training is specified, then measures of productivity, farm yield, appropriately correcting for other factors is required. Criteria, outcomes, and objectives are tied together in the planning process.

Step five is to summarize the major alternatives. If issues are stated in narrow concrete terms, they may serve to outline the alternatives. Otherwise, alternatives are often programs which could be employed to attain the stated objectives and which are to be judged for effectiveness by the criteria specified in step four. Alternatives in combating illiteracy include increased use of existing educational facilities but enlarging the scale and coverage, adult literacy programs, programs aimed at rural poverty and literacy such as Work Oriented Adult Literacy (WOALP) and mobile units aimed at short term instruction of the population. All of these programs will reduce illiteracy according to simple criterion-referenced tests; some will also work toward other objectives as well--such as direct productivity consequences of

instruction; and, finally, the time factor and coverage will differ among the alternatives.

Step six involves estimating the costs associated with each alternative. Analysts have to determine which processes of the alternatives give rise to costs, the categories of costs (i.e., fixed or variable), and the indirect cost consequences of the program both for individuals and society. Much of this information is technical in nature and may be well developed and available though this will probably be less true of the new or more innovative programs.

Step seven centers on assessing the effectiveness of each proposed alternative so that both costs and efficiency can be examined. This step ties in with the criteria adopted and the objectives specified to reach some conclusions about the efficacy of the programs. Each option has to be examined for its probable consequences. The timing and coverage of the consequences must be detailed as concretely as possible. And the assumptions involved in reaching conclusions must be made explicit. If simple, standard literacy grades are the criteria of effectiveness, then the probable impact on nationwide literacy of increasing school system coverage may be extrapolated on the basis of recent experience. Similarly, the rural impact of WOALP may be analyzed in terms of numbers who can be expected to pass standard literacy tests. In both cases the timing of the impact will differ and is one factor to include. In terms of the relevance of the assumptions of the analysis, the feasibility of, say, a program aimed at six week literacy training for inaccessible rural areas or differing language groups would have to be questioned. Simple extrapolation of current programs to

areas with widely differing characteristics may be extremely inaccurate.

As a consequence of assessing the effectiveness of the options, the planner becomes aware of obstacles to implementation, consequences of different programs, and additional data needs. Hence, step eight is simply a sorting out of the alternatives according to legal and administrative obstacles and investigating the basic lack of information and hence data needs required before reaching final decisions.

Step nine is analyzing the possible results of the programs. This involves melding costs and effectiveness measures in an attempt to rank alternatives. At this point, the three basic ingredients--objectives, costs, effectiveness--are combined to aid the decisionmaking process. This step involves not only assessment of the extent to which final objectives are attained and the costs of doing so, but also the added gains toward the measurable objectives that can be made at each level of program option and the added costs that would be incurred.

Step ten is the logical result of the previous analysis: the required lines of action are specified and, to as great an extent as possible, the exact procedures are documented. This step should be the natural consequence of carefully working through the previous nine steps. For example, a decision may be made to increase the number of teachers, expand physical plant, and increase the appropriate supplies in an effort to combat illiteracy--in short, to work through the expansion of the existing system. If this is the decision then based on the analysis involved in reaching that decision, the practical steps involved can be detailed. In addition, probable cost figures and the

timing of the expansion should be possible thus leading to multi-year planning documents which can be integrated into the current five year development planning process.

## CHAPTER XX

### OUTCOMES AND EVALUATION

Many nations presently are engaged in improving their policy-making and program performance capabilities. Evaluation--or the appraisal of the effectiveness of programs--is a part of that endeavor. Distinguished from policy analysis in that it is after the fact, evaluation can provide important data to improve analysis of alternative policies. In drawing on existing experiences, evaluation provides a guide to the "best" options.

Evaluation can range narrowly from project monitoring through various stages up to a broad concern for overall policy impact. Overall policy impact is concerned with the general system in which a policy is carried out. When evaluation is directed so broadly, it is necessary to identify components of the general system and assess the contribution of each. Policy impact evaluation in education, for example, has identified and assessed the role of the family, peer groups, neighborhood, and formal, as well as less traditional, education. (The Coleman Study and its subsequent reviews are examples of broad policy impact assessment.) Another example of evaluation directed at an overall policy impact would be one addressed to the questions of whether educational expenditure is achieving the purposes for which vast amounts of money are being spent.

Policy impact evaluation is heavily dependent upon a multi-disciplinary body of knowledge. For instance, if the issue is learning, researchers would have to draw from many fields of study: learning theory, teaching processes, and relationships of school to family, to neighborhood, to play and to work. Such evaluation--as one might well expect--is difficult due to our poor understanding of the outcomes of behavioral relationships and interactions.

Differential effectiveness assessment represents another level of evaluation which addresses itself to such questions as: Are some strategies more effective than others in achieving the purposes sought? Or does a particular program work better or cost less for one group of persons than another?

A third level, project monitoring, examines specific activities rather than overall developmental programs. Much of the emphasis of project monitoring is on getting reliable data about the project both as to cost and activities.

However broadly or narrowly directed, evaluation remains essentially a process for measuring progress and answering questions such as: Is the activity, program, or policy a success? Those questions cannot be answered unless there are yardsticks available by which to assess progress. Without such "yardsticks of success" evaluation is impossible; with them, it is encouraged. Furthermore, it is not a simple matter to determine what the outcomes of any activity are. Most programs have multiple purposes. What then to evaluate is a hard issue, and what is more, not all identifiable outcomes can be evaluated, for some are more difficult to get data on, and in many cases evaluations are costly.

Nevertheless, there is an easier path. One can choose to evaluate only certain outcomes, namely, those outcomes that are sought from any activity. It is often difficult to choose among objectives and the measures of outcome appropriate to those objectives. Agreement on outcomes is likely to be difficult. A workable formulation of outcomes that specifies proxy measures can be achieved without reaching such a full accord. It is easier to reach this point when a number of indicators are used rather than one. For each, the question to be addressed is: What difference did the program make?

Designing evaluation studies in terms of selected outcomes is set forth below as a series of requirements taken from a U.S. draft report:<sup>1/</sup>

---Program evaluation studies should isolate and specify the various elements and phases of the program studied and the several elements and phases of its physical, biological, general social, organizational, and economic environments that affect its operation and their results, and to state fully the causal and other relationships among them.

---Functional relationships of benefit, effectiveness or output variables should be hypothesized in terms of resources and conditions and constraints imposed; the evaluation should seek to accept or reject these hypotheses and quantify them.

---Program models should be utilized when they can generate more useful, comprehensive, or economical results to the fullest extent feasible; the evaluator is to choose or devise the most relevant model

---

<sup>1/</sup> Draft report from the General Accounting Office.

and use it with rigor. Such models will depict the flow of resource inputs into activities that cause results as conditioned by environmental and institutional constraints.

---Statistical estimation procedures should be designed to quantify the variables and model inputs from feasible sources of known validity; if validity is in doubt and either time or resources do not permit new data gathering, the risk of erroneous conclusions should be weighed before proceeding.

---Sampling procedures should conform to recognized published standards that can be documented and referenced or the departure from such standards should be explicitly and clearly explained and justified.

---Assumptions used at any stage of the inquiry and in all constituent analyses of the evaluation study should be realistic, clearly specified, and to the extent possible, verified as to accuracy. Any recognized bias in the assumptions should be identified and clearly stated. The evaluator should continuously reassess assumptions with which the study started and change them and recycle the analysis when this appears needed.

---The assessment should be sufficiently broad in scope as to provide findings that can satisfy lines of inquiry required to understand the results of the educational services.

One more point should be added at this time about the design of evaluation studies. Evaluations tend often to be static in their design. While structuring a system to study and define output measures for current programs, it is important to build into that structure the capability for change and evolution. As more becomes known, new program goals and objectives will be defined and old ones abandoned.

Changes may not come in large quantum jumps, but rather in small re-alignments, resulting from new knowledge, new perceptions, and a broader understanding of the problem and its potential solutions. Evaluation capabilities must improve and be redeveloped over time as new and different education programs enter the picture. We must view evaluation as a kinetic, dynamic function rather than as the application of old, static, or perhaps obsolete criteria to assess programs that are constantly developing and changing.

Once the outcomes to be evaluated are selected and once the evaluation program is designed, one must begin collecting data. This reservoir of data should be as complete as necessary to support a full analysis. Existing records--maintained so as to be auditable--; surveys of target groups, administrators, or others interested; periodic reports--referenced to records or other sources--; and other research and program evaluation study reports should all be utilized. Ideally, the data should reach all parameters of the program, including side-effects, by-products, and secondary results. But, realistically, this is often impossible, for much of the data collected is likely to be proxy, that is, partial or symptomatic information standing in for a general category of results more difficult to gain complete data on. For example, in assessment of earnings in the computation of investment returns, a proxy--"average earnings"--often stands in for marginal wages; and wages at the margin may be for all classes of workers rather than workers with the specific characteristics of those under study.

But, in any event, whether "proxy" or "full-parameter" information is used, only those statistical procedures that are well known and

readily interpreted should be utilized for collecting and analyzing any quantified information. Any uncommon statistical techniques employed should be clearly referenced to authorities who developed or tested them.

Another important part of evaluation is project monitoring. As indicated earlier, it is mainly an administrative function, but inter-project comparisons can yield experimental findings. The extent to which such findings can be drawn from the variation among projects depends upon objective observation and the capacity to find true controls. That is to say when we seek an answer to the question, what difference does the project make, we have to ask: Differences compared to what? And the basis for comparison must be some project that provides the function of an experimental type of control.

Indeed, it has been advocated that planned variations be built into projects to serve as control groups for the purposes of experimentation. In any event, the normal process of attempting to judge "success" generates experimentation, for in the process of so doing it becomes clear that much necessary knowledge is lacking about program development and the way in which outcomes can be obtained. While experimentation in social programs is difficult, evaluation moves in that direction because of difficulties of telling whether or not programs have had the intended effects.

There are, then, minimal requirements of a comprehensive project evaluation, if it is to provide useful policy guidance over the program: (a) the evaluation system must provide for autonomy of evaluation efforts; and (b) evaluation systems must provide evaluators

with compliance control.

The vested interests, biases, prejudices, short-sightedness and honest convictions of those persons closely associated with programs may hamper objective evaluation if they are involved in the evaluation process. Administrators and staff establish biases that severely limit their ability as a group to contribute to evaluation efforts. For this reason, there is good cause to be somewhat skeptical of data or in-house evaluations provided by persons working within specific programs.

Evaluation--like all other human activities--is constrained by limited finances. Independent investigations--though ideal--cost more in terms of time and personnel. Evaluation must, therefore, depend partly--often largely--on those persons working within the project to provide essential information and perspective. Given the incentives noted above for persons working in particular programs or projects to bias or block evaluation efforts, the agency charged with evaluation must have the legislative authority to obtain information. However, compulsory compliance forced upon project administrators and staff will probably not yield useful results in terms of subsequent action. Half-hearted or counterproductive efforts forced out of program or project staff is not the goal of compliance control. The knowledge that deliberate lack of a cooperative effort may result in an evaluation report biased against the program should be incentive for program personnel to provide necessary data. The goal of the evaluating agency should be a cooperative effort with a substantial and fair hearing of the ideas and opinions of those within the program. This all suggests that the structure of the evaluation function must be hierarchical,

involving persons from the project level on up with adequate shares of program data accessibility and evaluation objectivity.

### The Uses of Evaluation

Evaluation has a number of uses:

1. Defining program potential--Evaluation of past performances and goals often deflates unwarranted optimism and gives planners a better idea of what is feasible within certain resource constraints.

When this consequence of outcome measurements is fully reflected in policy decision, it will be clear that the generalities of earlier days have been abandoned and in their place there is far greater specificity borne of the greater knowledge acquired of program operations in terms of measures of "success" and "results." For example, through evaluation of education, we have defined better the role of the family, neighborhood, and peer groups in learning, and consequently have obtained more realistic views about what can be accomplished during the limited time spent in school.

2. Targeting resources--Outcome measures facilitate a focus of policy on critical needs. When outcome data questions disparities in allocation of resources, new emphasis is given to marshaling those resources where they can meet the most pressing need more effectively.

3. Encouraging program change--Outcome data based on established criteria, moreover, makes it possible to deal with activities and programs in a functioning and changing system. The assessment of results encourages monitoring of projects to determine whether they are on target; it also paves the way for subsequent changes in direction by revealing weaknesses and barriers.

4. Facilitating cost comparisons--The criteria--once available in measurable form--permit analysis of costs per unit for a defined result. Unless program output can be equated from one activity to another--or one program to another--cost comparisons are not very meaningful. Standardization of "units" are essential to such cost comparisons.

5. Better understanding of program interaction--Derived from outcome assessment too is a better awareness of the relationships of programs and activities to human behavior. Evaluation also brings a better understanding of the problem of duration and timing in public programs and activities. For instance, we have begun to give more emphasis to the length of time required to produce a given output with the resource available. And, for learning, the question of optimum timing is becoming more vital.

6. Encouraging a "numbers game"--Measurements that are incorporated in resource and program decision lend themselves to abuse. The more important the numbers become, the greater the incentive for abuse. Aside from inaccurate reporting--which may be of less importance--one of the aspects of such abuse of the statistics from evaluation of program policies are focused to yield the most favorable number result. If grade-repeaters in school are an index of a problem in teaching, and the fewer repeaters the better, automatic promotion of even poor students becomes a policy. Or if the number of graduates is deemed the measure of success, graduation of nearly all is the response. The changes in numbers alter the criteria count but not the knowledge or skill of young persons in the nation.

Measurements of the objectives of any program tend to take on a dynamics of their own to which policy officials respond, so it is

easy to understand now the abuse of statistics and evaluation tends to alter the very objectives themselves. Pressures to succeed, by whatever criteria, should not be discounted and certainly not ignored.

7. Evaluation as an educational tool--We earlier reviewed briefly outcome measures as educational devices. Evaluation using outcome criteria is useful as a guide to teaching, curriculum formulation, and to administrators who carry responsibility for assuring that the criteria used are understood and reflective of teaching needs. In the case of programmed instruction, detailed outcome objectives are fundamental to the design of the programs. In some instances, individual student learning on a tutorial basis is formulated in terms of essentially a contract between student and teacher. Rating of progress toward specified goals is an integral part of this educational method.

Before concluding, it is important to point out at least three warnings for the use of evaluation studies.

First, while evaluation with its requirements for specificity of criteria of assessment offers great promise, it lends itself to premature political conclusions and manipulations. Politicians may draw too hastily upon evaluation findings.

Second, beginning efforts at evaluation have to be viewed as beginnings. One evaluation study, no matter how well carried out, is not a sufficient guide for a major policy decision.

Third, learn when not to use evaluation's findings. A dean of the evaluation profession recently stated in discussion of the negative findings of the Westinghouse-Ohio University evaluation of Head Start, "In this case I think that the testimony of the parents who had children in school earlier and then had children in Head Start

programs--their warm enthusiasms--were more valid indicators of the impact of Head Start than the stacks of computer output. I am glad that the political system disregarded a very impressive statistical analysis."