DOCUMENT RESUME

ED 041 936 TM 000 028

AUTHOR Evans, David
TITLE International Education From a New Quantitative Perspective.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
PUB DATE Mar 70
NOTE 16p.; From symposium "Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques," (Annual meeting American Educational Research Association, Minneapolis, Minn., March 1970)

EDRS PRICE EDRS Price MF-$0.25 HC-$0.90
DESCRIPTORS Academic Achievement, Competitive Selection, *Curriculum Evaluation, *Developing Nations, Educational Change, Educational Improvement, Educational Innovation, Educational Planning, Educational Policy, *International Education, Item Sampling, Symposia
IDENTIFIERS *Comprehensive Achievement Monitoring

ABSTRACT

Educational testing and measurement on an international basis is viewed from the new quantitative perspective provided by the Comprehensive Achievement Monitoring (CAM) method. The technique of longitudinal testing through item sampling employed in CAM is seen as being of significant value in formulating a systematic measure of achievement across national boundaries. Some specific applications of the principles embodied in CAM to the international educational situation are discussed. These include student selection, curriculum and program evaluation, and system monitoring and quality control. These applications are seen as particularly beneficial to developing countries which are attempting educational development and reform. (PR)

INTERNATIONAL EDUCATION FROM A NEW QUANTITATIVE PERSPECTIVE [1]

by

David Evans
School of Education
University of Massachusetts

[1] A paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, 1970, in a symposium entitled " Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques ", sponsored by the Project for Comprehensive Achievement Monitoring, School of Education, University of Massachusetts, Amherst, Massachusetts.

INTERNATIONAL EDUCATION FROM A NEW QUANTITATIVE PERSPECTIVE

David R. Evans
School of Education
University of Massachusetts

The emerging interest in educational testing and measurement on an international basis can be traced to at least two major historical trends: one growing out of the field of comparative education, aided by the recent efforts of UNESCO and other international bodies, and the other deriving from the transfer of testing methods as part of the educational system which metropolitan countries set up in their colonies in Africa and Asia. As these two trends grow and merge they have begun to stimulate a search for new and more efficient testing procedures which, coupled with the power and flexibility available with modern computers, can be used to measure, compare, select, and improve education on an international basis. This paper will review briefly the current situation in terms of these two trends and will then look at potential applications of some new measurement procedures, particularly in the developing countries of the world.

Until the latter part of the 1950's, comparative education consisted primarily of the juxtaposition of analytic case studies of educational systems in different countries. Authors were content to discuss systems in a qualitative manner relating education to characteristics of the society and discussing the interaction between educational styles and national character (see, for instance, Noah & Eckstein, 1969). With the Russian success in launching the first satellite and the subsequent critical analysis of Russian and American educational systems, there emerged an emphasis on quantitative data. Data appeared which reflected enrollment ratios, proportions of students in science and humanities, and even attempts to quantify the quality of educational output of various countries.

Supporting this trend were the growing efforts of international organizations like UNESCO and OECD to collect and publish comparable data on education in the member countries of their organizations. An even stronger demand for quantitative data was produced by the burgeoning growth of national planning efforts in the 1960's. Early attempts at national educational planning quickly revealed the need for extensive and accurate statistics of a kind which had never before been collected. The result was a rapid growth in data on a wide range of inputs to the educational system, information on processing characteristics of the system, and studies of output measures in relation to national economic statistics.

Attempts to construct models of the relationship between national educational and economic systems soon demonstrated the need for measures of output that reflected quality and content rather than just aggregates of numbers at various educational levels. At the same time, educational researchers were beginning to tackle the problem of comparing the amount of learning produced atdifferent levels in different countries. After early efforts comparing two groups within a country or two countries (for example, Kramer, 1959), and more extensive pilot projects in a dozen countries (see Fodhay, 1962), a major international testing project in mathematics was planned and carried out in twelve countries.

The results of this project are reported in a two volume work edited by Torsten Husen (1967). The project involved testing two basic school groups in each country using a battery of nine tests from which three or four were selected depending on the level being tested. The contents of the tests were divided into twelve mathematics skills categories (Husen, 1967, Vol. I, pp. 104-105). Each country responded to the whole range of mathematics competencies. The test was administered once to each pupil in the sample and each pupil at a given level responded to the same items regardless of country. In addition, data was collected on schools, teachers, curriculum and the national educational system in each country. The data was then used to test various hypotheses relating these independent variables to performance in the defined mathematics categories.

The other major trend in international testing, educational testing for selection, does not have any comparative goals, although the information produced would be very useful if analyzed on a comparative basis. Examination syndicates in the metropolitan countries still set and mark selection examinations for many of their former colonies. The largest operations are carried on by the Cambridge and Oxford syndicates in England. Comparable services are provided by a section of the French Ministry of Education to the franc zone. Since the advent of independence most of these territories have gradually moved toward localization of the administration of these exams, often on a regional basis as exemplified by the regional examination boards in West and East Africa. However, even in these areas the exams retain their former structure and are still frequently marked in England.

The result of a common colonial history, and therefore a common design in educational systems, is a remarkable uniformity in the style and content of education throughout English speaking Africa and to a lesser extent Asia. In terms of testing, the systems take the form of a sequence of selection nets of ever decreasing mesh size. In most countries there are currently three major examination hurdles which determine the occupational destiny, and to a large extent the future life style, of the pupils in school. These occur at the end of primary school (now seven years in many countries), at the end of four years of secondary school (sometimes known as "O" level), and at the end of two years of Higher School. Figure 1 indicates the pattern which is present in Uganda and which if fairly typical of English speaking tropical Africa.

Because of the acute shortage of places at the higher levels, the examination is extremely selective. Examinations at each level consist of a battery of tests in the standard school subjects, usually with two or three papers in each subject. Exams are given at the end of the terminal school year for each level and generally take a week or more to administer. Tests are given on a national basis with every pupil taking exactly the same test at the same time throughout the country. Selection for the next level of schooling depends almost exclusively on some sort of cumulative mark computed from the various subject exams. Only in cases of equal marks where borderline

candidates must be selected do other factors such as age or character references influence the decision (see discussion in Somerset, 1968, p. 17ff.).

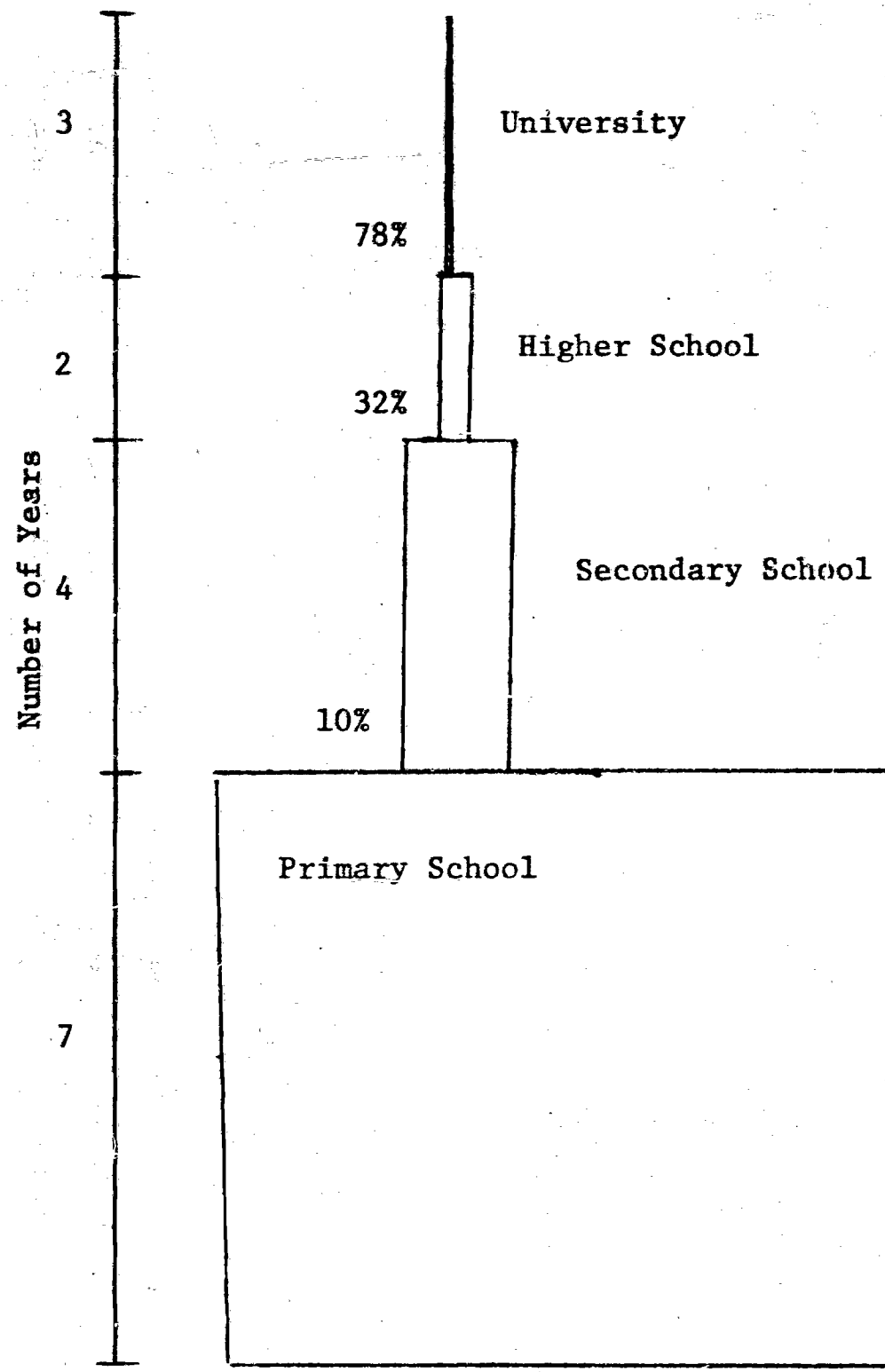
The importance of the examination results to the pupils cannot be easily understood by the Western observer. Failure to enter secondary school, and only one in ten of those who successfully complete primary school do, virtually assures the pupil that he must remain in a rural area or scramble for a low paying job in the growing urban slum areas. Access to a status job, enabling the individual to reap the benefits of modern technology, is, for all purposes, closed. As a result, the examinations have a tremendous psychological impact on the pupils and have an all-pervasive effect on the content and teaching goals of the school system.

In terms of reliability and validity, the test format and the extreme pressures under which the candidates work combine to suggest that there is considerable room for improvement in the whole testing process. A recent study, by Somerset in Uganda, of the predictive validity of the Primary Leaving Examination in terms of subsequent performance on the School Certificate Examination found that correlations between the two exams were of the order of 0.4 or lower. Regression analysis showed that the predictive ability for students with borderline scores was a good deal lower (Somerset, 1968). With extremely scarce resources, developing countries are in need of efficient testing and selection systems which insure maximum utilization of ability.

Principles Underlying New Approaches to Achievement Testing

A new approach to testing which has important international implications is called Comprehensive Achievement Monitoring. The system is longitudinal, uses item sampling, ties each item to a specific objective, and is based on a completely specified set of objectives imbedded in the teaching goals of a course of study. Continuous information is collected on each objective prior to its being taught, immediately after it is taught, and periodically afterwards to measure retention. Testing takes place at regular intervals throughout the course of study by means of a set of parallel tests of comparable

Education Profile for Uganda



Third Level

Asia 2.0 (0.5 - 5.7)
South America 3.5 (0.8 - 10.0)
Africa 0.14 (.01 - .53)

Second Level

Asia 16.4 (7.3 - 30.9)
South America 16.0 (5.5 - 32.0)
Africa 4.0 (0.3 - 21.9)

First Level

Asia 42.3 (20 - 63)
South America 52.6 (19 - 69)
Africa 24.3 (3 - 44)

Education Enrollment Ratios: Avg. (range)

Source: Harbison & Myers, 1964,45-48.

FIGURE 1

content and difficulty. All forms of the test are administered at each testing point so that information is available on all the objectives of the course at each testing period. Because all the test forms are used each time, each form is relatively short and a given pupils gets a different test each time (for more details see Wightman & Gorth, 1969).

The most direct product of this method is the ability to trace achievement on each objective throughout the course to see what the effect of teaching was, to see how one objective interacts with the teaching of another, to see when pupils actually learn, and to follow their retention patterns. Depending on the goals of the user, the results can become a major teaching device as students monitor their own progress, the system can be used primarily as a selection process, or as a system monitoring technique to measure national achievement, or to provide the ministry with accurate quality control measures for different parts of the country. A number of specific applications of the principles embodied in Comprehensive Achievement Monitoring will be pointed out in the following section.

Cross-National Achievement Testing

The study of mathematics achievement in twelve countries referred to in the introduction represents the beginning of systematic attempts to measure achievement across national boundaries. There are a number of ways in which the design utilized could be modified to increase efficiency by making use of some of the principles outlined above.

For example, a significant imrpovement in reliability and validity could be achieved by going to an item sampling procedure so that each national population would respond to a number of items keyed to a desired objective rather than having the whole population respond to the same one or two questions for that objective. This procedure is much more efficient in that each pupil still answers only one or two questions, but from the whole population you cane make group estimates based on as many as twenty or thirty questions for a given objective. The only extra costs appear in the need to code a number of parallel forms of the test and the need for a somewhat more sophisticated

data processing and storage technique. The rate of improvement in the capacity and capability of computers makes the latter problem relatively easy to solve. Such programs already exist for use on smaller sample (see, for instance, Gorth, Grayson, & Stroud, 1969).

A logical extension of this sampling process could be used to cut the length of time an individual student spends taking a test. By sampling objectives as well as items for a given groups of students, each student only answers items relating to a sub-set of the objectives on which the whole group is being measured. A double benefit would result: individual testing time would be cut and the range of objectives tested could probably be increased for most groups. On the other hand, such a method would require that student groups be defined ahead of time in terms of the independent variables relating to school type, rural/urban setting, school quality, etc. The twelve country mathematics study could be useful to indicate those variables which are most likely to have an effect and thus should be controlled for in future projects.

Of necessity, the original hypotheses in the Husen study are painted with fairly wide brush strokes. However, with that data to point the way, more specific hypotheses can be considered, particularly as they relate to more finite areas in mathematics. A desirable adjunct to more specific hypotheses would be the keying of all items to precisely defined objectives. Objectives could then be grouped along dimensions, other than the standard sub-categories, to test hypotheses relating to certain skills rather than content. Other possibilities arise and are limited only by the extent to which one can calibrate items along different dimensions. The result of linking items to specific objectives is a much greater flexibility in the type of analysis which can be undertaken; one is no longer limited to aggregate scores in traditional categories. In addition, hypotheses linked to specific objectives are more likely to have direct implications for local and national school policies.

For certain kinds of cross-national testing, one might also consider going to a limited longitudinal design, perhaps to smooth out differences in

school years or differences in syllabus design. For instance, pupils could be tested three or four times over an academic year. Using both item and objective sampling, and administering all forms of the test at each point would mean a relatively short test for any given pupil. Under certain circumstances one could also relate the time when various objectives were taught to the testing time in order to study the relationship between time of presentation and performance on tests. Such information would be particularly relevant when one is interested in the effects of different curriculum, teaching methods, texts, and other materials over time.

Testing in Developing Countries

While the techniques of Comprehensive Achievement Monitoring are readily transferable to the educational systems of the more developed countries, their application in the developing world requires some adaptation. Three areas of application are worth considering: student selection, curriculum and project evaluation, and system monitoring and quality control for educational planning.

Student Selection. As indicated in the introduction, testing in developing countries is now confined almost exclusively to a sequence of test batteries administered at major breaks in the system for the purpose of selecting those few who will go on to the next level. Testing done between these major selection points is either done as a training procedure to get students prepared for the selection exams or for the purpose of ranking pupils in the class at the end of the year. As there is no systematic attempt to keep cumulative records on individual pupils, a student's school record is generally confined to his cumulative score on the highest level selection examination which he took.

The problems with the selection examinations as they now stand are related to their one-time, all-or-nothing nature, and the limitations imposed by the need to do all the testing within a short time interval. Validity and reliability could be considerably improved by going to some type of longitudinal measuring system, perhaps during the last six months prior to the selection decision.

If the objectives for the examination were completely specified, and a group of items written for each objective, then a series of monitors of parallel form could be constructed. The length of the monitors would depend on the overall number of objectives and the number of testing periods. Several hundred objectives grouped into twenty categories and spread out over ten test periods would lead to tests of twenty to forty items each, depending on decisions about the number of items desired for each detailed objective.

Since the curriculum in developing countries tends to be more cyclical than sequential, most of the items would be of a post-test nature. Patterns of review during the six month period would vary across schools and teachers, but the fact that all forms of the test are used each time would compensate for any temporary advantages to a particular pupil as the result of one testing. One might object that this would lead to the last six months being spent entirely on review and testing; this is the case currently. The advantage of the longitudinal approach lies in the lack of psychological pressure and the increased reliability from repeated testing. The tests also form a continuous monitoring system which help the pupil evaluate his progress and to learn more thoroughly the competencies being tested because of the periodic reinforcement of the whole range of skills.

Selection could then be made on the basis of a pool of 200 to 400 items covering the complete range of objectives. By grouping objectives, sub-scores could be computed to reflect specific skill clusters which could then be weighted to bias selection in desired directions according to manpower needs of the country. A longitudinal testing scheme also makes selection possible on the basis of information on the rate of learning of pupils over the testing period. Rate of learning information along with measures of initial and final performance might well be used to offset the severe handicaps imposed on children from poor quality schools. Somerset concludes from his study of education in Uganda that the quality of elementary schooling has an irreversible effect on secondary school performance (Somerset, 1968, p. 97). The review and reinforcement nature of the testing sequence might offset part of this difference over a six month period.

Problems with this system might be encountered in terms of the increased communication and liaison required since test results must be returned ten different times. On the other hand, the very frequency of the event should lead to rapid refinements of the system so that it could operate fairly efficiently. The other major problem would be one of security; test booklets would have to be kept secure in each school over the testing period. Several alternatives might be possible, such as having them kept in a regionally centralized location between testings. The risk is also lower since the cost of a leak for a given test form would be relatively low since it comprises only a small part of the total item pool.

Because of the number of unknowns in the proposed approach, the author would recommend that the system be used with a sample of students in parallel with the normal system. In that way results of the two could be compared and data could be acquired on the relative costs and benefits of the two systems. Spreading the testing over the longer period would also reduce the peak load demands on processing equipment and personnel, allowing a less intense and more regular utilization of these resources.

Curriculum and Project Evaluation. A second area of application which has considerable potential for developing countries is that of curriculum and project evaluation. This is particularly appropriate at a time when national and regional institutions are engaged in rewriting syllabi and curriculum to adapt their inherited system to the specific needs and interests of African pupils. Curriculum evaluation is perhaps the weakest aspect of educational systems in developing areas. What little evaluation is done is usually of a vague and general nature. Typically, the only criteria is performance on the selection examination.. Sometimes the examination is not even vaguely related to the syllabus to be evaluated, and it may occur as much as four years after the particular curriculum being evaluated. Unfortunately, the social and political pressures to consider the selection examination the only valid criterion are extremely strong. The result has been that most new curriculum projects have enjoyed brief popularity and then faded under the relentless pressure of the selection examinations.

Compounding this difficulty is the extreme scarcity of personnel with testing and evaluation skills. Only rarely does a ministry inspectorate contain anyone with testing knowledge. Examination secretaries are primarily administrative positions and rarely have any training in testing methodology. Added to this is the fact that the teaching force in primary school is in need of massive retraining in subject skills included in the curriculum, to say nothing of measurement procedures. Secondary school teachers are well trained in their subject areas, but for a variety of historical reasons they have little knowledge of testing and measurement skills.

Such a situation seems ripe for curriculum imbedded testing procedures. This means that when a new curriculum is constructed a complete set of objectives is specified at the same time. Items are written to measure each of these objectives, and a set of parallel form monitors is constructed for inclusion in the teaching materials for the curriculum. Teachers are instructed in the use of the monitors and all results are forwarded to evaluators on a regular basis. At the same time, copies of results can be returned to students for their own information and motivation.

One advantage of using curriculum imbedded testing is that in order to construct the tests, curriculum planners are forced to specify a detailed set of objectives from which to work. General objectives are included in most projects, but often they are only partially complete since they are not directly applied in the traditional curriculum. A second advantage of this approach is that test results become a direct measure of these specific objectives. As a result, evaluators have information with which they can plot achievement profiles for each of the major goals of the project and can make modifications on the basis of specific data about the success or failure of each aspect of the project. In contrast, traditional methods of evaluation involve all-or-nothing decisions based on performance on tests which are only indirectly related to the goals of the new curriculum.

Another advantage lies in the longitudinal format of the testing. Because each objective is tested before it is taught, immediately after it is taught, and regularly thereafter for retention, achievement profiles for each objective can be plotted to trace the relationship between achievement and the teaching

of that objective and also the teaching of other related objectives which may influence the first objective. Such data opens the way for decisions on the sequence of teaching objectives, the length of time spent on various objectives, and the interation of teaching methods with the abilities of different groups of students. Finally, continuous data allows curriculum designers to modify the approach as the course goes along rather than having to wait until the end of the year for any measure of success or failure.

Evaluating curriculum with Comprehensive Achievement Monitoring provides decision makers with a range and specificity of data which greatly increases the alternatives available to them when they must decide on the future of a new curriculum. Instead of the traditional dichotomous decision of whether to drop or retain a curriculum, the decision maker can now use the data on the strengths and weaknesses of specific parts of the curriculum to recommend a range of deletions, improvements, and modifications. The data also provides information on the relative efficiency of the methods employed to teach the curriculum in terms of learning speed and retention.

When a particular curriculum is subsequently adopted, the imbedded tests used for evaluation can become part of the teacher's resources for teaching and monitoring. This is especially useful when teachers are poorly trained and have neither the skills nor the motivation to construct good quality monitors. The imbedded monitors thus improve the measurement of pupils' skills and, at the same time, can be used as a teaching device for pupils to get regular feedback on their progress which enables them to plan their study time. Not incidentally, such monitors would also serve as models for testing which teachers might want to imitate in their other courses.

System Monitoring and Quality Control. A third possible application of CAM techniques would provide a completely new type of information for the inspectorate and for ministries of planning and development. Since the syllabi for all subjects are set and controlled on a national level, it is feasible to consider monitoring progress in specific subjects on a national level. The information produced would serve such goals as quality control, input on the efficiency and productivity of different institutions, input

for decisions on resource allocation and educational policies, and badly needed data for constructing educational plans for future expansion and investment in education.

At present the only quality measures available to the inspectorate are the wholistic results of selection examinations given at the end of each major level in the system and the subjective impressions of the inspectors as they visit schools and watch individual classes. To provide the needed data, a national sample of students could be selected in a particular subject and, over a period of as much as an academic year, administered a sequence of CAM monitors based on a set of nationally specified objectives. There need only be one set of monitors so that test construction is a task of reasonable dimensions. Participating teachers can hand score results before sending the data to a central location. With a short training session, teachers could then make use of the data as part of their teaching. Monitors can be short so that they are easily administered within a period on the normal school schedule. Past experience indicates that teachers and pupils quickly adapt to the system and often come to view the monitors as desirable parts of the learning process rather than as tests to be feared.

The information available to the inspectorate and the ministry then changes radically. There would be information on achievement on the complete range of specific objectives in a given subject curriculum. Depending on the choice of the sample, the achievement can be related to type of school, year in school, quality of pupils and characteristics of teachers. Achievement profiles can be constructed by region, by type of school, or by virtually any other delineated variable. In addition to measures of achievement, the longitudinal nature of the monitoring provides information on rates of learning and retention, interaction with teaching methods and materials, and interaction between objectives. Again, these relationships can be studied within the context of other variables such as pupil characteristics or teacher experience.

With a monitoring program extending over several years, the inspectorate could build a composite picture of learning and retention rates across the nation as students moved through a four year curriculum in mathematics, for

example. The potential uses of such information are of a wide variety. Among them would be decisions on the use of manpower in the inspectorate, choice of teacher populations for in-service training, choice of content for such training, revision of curriculum, particularly in terms of sequence and time allocations for the different sections, and larger policy decisions about the length of time spent at various levels in the educational system.

The last example indicates ways in which the data would influence national educational policies beyond the question of quality control. A good illustration is provided by the decision several years ago to reduce primary education in Uganda by one year, to a total of seven years. The decision was taken on the basis of extensive information about the cost of the final year and on the basis of the political need to provide more school places at the same level of financing. The crucial variable, a measure of the productivity of that year, was completely missing. There was a general intuitive feeling among educators, backed up by sporadic testing information, that the marginal productivity of the last year was not commensurate with its cost. In contrast, consider the detail which a CAM process would have provided as input to such a decision. Even more important, the data would provide guidelines as to where the curriculum needed to be modified in order to excise a year of classes.

When combined with detailed information on inputs to the school system, the CAM data on specific outputs will enable planners to relate particular inputs to particular changes in outputs. This leads to measures of efficiency of various types of educational institutions which in turn become input for decisions on allocation of resources for maximum effectiveness. Questions on the use of various types of teachers, for instance, can be related to output effectiveness of teachers. Such data could be used to investigate charges that young volunteer teachers from the United States are less effective and produce lower examination results. Another application would be to understand the relative strengths and weaknesses of the different grades of teachers in the system. Many systems in Africa are now producing a lower grade of teacher to be used in the first part of secondary education. How should these teachers be used to be most effective? Parallel decisions

are necessary in areas of finance, capital expenditure, and pupil selection policies, and all would benefit from specific, time-based measures of output for comparative valuation of different inputs.

Finally, both the short and long term results of a national monitoring program would be invaluable input for educational planners who must project investments in education for five and ten year periods. At the moment they work with only the crudest measures of outputs and their valuation. Being able to disaggregate educational indices so as to include measures of specific skills and competencies would be a considerable improvement over current methods which depend primarily on number of years of schooling completed. Long term results would provide much more useful measures of the stock of educated manpower. Regular monitoring over a period of five to ten years would produce accurate profiles of stocks in terms of very specific skill achievements. This data would in turn allow an attack on the serious problem which manpower experts now face in translating occupational categories into educational prerequisites.

The kinds of benefits cited in the last few paragraphs are particularly important for developing countries. Their educational systems are relatively small, fairly uniform, especially at the second and third levels, and they are administered on a centralized national basis. Coupled with these characteristics are the facts of extreme resource scarcity in the face of demand levels several times beyond the capacity which their resources can hope to support. In short, efficient planning and allocation of resources is imperative both economically and often politically. In such a situation the use of efficient monitoring procedures has a much more favorable cost-benefit position than it may have in more developed countries.

REFERENCES

Foshay, A.W. (ed.) Educational Achievements of Thirteen-Year-Olds in Twelve Countries. Hamburg: UNESCO Institute for Education, 1962.

Gorth, W., Grayson, & Stroud. "A Computer Program to Tabulate and Plot Achievement Profiles of Longitudinal Achievement Testing Using Item Sampling," Educational And Psychological Measurement, 1969, 29, 179-180.

Harbison, F., & Myers, C. Education, Manpower and Economic Growth. New York: McGraw-Hill Book Company, 1964.

Husen, T. (ed.) International Study of Achievement in Mathematics: A Comparison of Twelve Countries. Two Volumes. New York: John Wiley and Sons, International Project for the Evaluation of Educational Achievement, 1967.

Kramer, K. "Arithmetic Achievement in Iowa and the Netherlands," Elementary School Journal, 1959, 59.

Noah, H.J., & Eckstein, M.A. Toward a Science of Comparative Education. London: The Macmillan Company, 1969.

Somerset, H.C.A. Predicting Success in School Certificate. Nairobi: East African Publishing House, 1968.

Wightman, L., & Gorth, W."CAM: The New Look in Classroom Testing," Trend, 1969, Spring, 56-57.