DOCUMENT RESUME

ED 041 934 TM 000 024

AUTHOR Allen, D. W.; and Others
TITLE An Introduction to Longitudinal Testing Using Item Sampling Techniques: Comprehensive Achievement Monitoring.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
PUB DATE Mar 70
NOTE 21p.; From symposium "Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques." (Annual meeting of the American Educational Research Association, Minneapolis, Minn. March 1970)

EDRS PRICE EDRS Price MF-$0.25 HC-$1.15
DESCRIPTORS Academic Achievement, Course Objectives, Evaluation Methods, *Evaluation Techniques, *Item Sampling, Student Development, Symposia, Test Construction, *Testing
IDENTIFIERS *Comprehensive Achievement Monitoring

ABSTRACT

Comprehensive Achievement Monitoring (CAM), a new evaluation method, systematically and comprehensively measures student achievement. The unique characteristics of this model are described in detail, as are those of two more familiar models, classroom testing, and curriculum program evaluation. The amount and quality of information available from each model is summarized. To obtain the goals it sets for reliability and validity, CAM employs two modern measurement techniques, item sampling and longitudinal testing. It tests achievement on every objective of a course at frequent test administrations throughout that course. CAM yields more information of superior quality than do the other models, due to systematic pretests and measures of retention of objectives, and also allows for comprehensive feedback. This systematic accumulation of data is of crucial importance in evaluating education on a state-wide basis. (PR)

An Introduction to

Longitudinal Testing Using Item Sampling Techniques:

Comprehensive Achievement Monitoring

By

D.W. Allen, W.P. Gorth, and L.E. Wightman

School of Education, University of Massachusetts

Amherst, Massachusetts 01002

Designing instructional systems requires detailed information about the impact of alternative instructional treatments on the learning of the students. With the constantly changing forms of curricula new methods for evaluating student achievement were needed. The Charles F. Kettering Foundation in a grant to the principal investigator, Dwight W. Allen, and under the project director, William F. Gorth, both of the University of Massachusetts, has supported the search for and the development of such a new evaluation method, Comprehensive Achievement Monitoring, CAM.

CAM measures achievement in a systematic way throughout a course in the secondary or elementary school. It is comprehensive in two dimensions: 1. Time because achievement is measured throughout a course and 2. Course content because achievement is measured on all of the behavioral objectives specified for a course at each time. CAM uses several of the most modern techniques in educational measurement to obtain the goals it sets for reliability and validity. The techniques include item sampling which has been recently been devloped by Fredrick Lord and longitudinal testing which has often been reccommended to measure change or growth. Both of these ideas have been tied to computer programs for rapid analysis and reporting of the results to students, teachers, and administrators.

A Comparison of Three Evaluation Models

Comprehensive Achievement Monitoring (CAM) will be described and its unique features characterized. After this description two more familiar evaluation models (e.g., usual classroom testing, and curriculum project evaluation) will also be described and their strengths and limitations mentioned to contrast them with CAM. Finally, the pattern and quality of information generated by the models will be compared.

Description of Comprehensive Achievement Monitoring

CAM is a system for testing achievement on every objective of a course, at frequent test administrations throughout the course. At each test administration, performance on objectives not yet taught is pretested, performance on objectives just taught is immediately post-tested, and performance on objectives taught earlier in the course is measured for retention. Parallel test forms, comparable in difficulty and content, are all used at each test administration, but each student receives a particular form only once during the course. Each form typically has an item for each objective. Each item is used on only one test form. The function of a particular item changes in relation to the time at which its objective is taught. Testing may take place at regular intervals (e.g., every two weeks) or at the end of certain instructional units. Computer based analyses and reports are made.

Specification of objectives. The most fundamental preparatory step for the use of CAM is the specification of the objectives to be evaluated, in testable, behavioral terms. Objectives may be categorized according to numerous dimensions, and possibly organized into instructional units. Written objectives for a variety of closely related projects or courses may be collated and pooled. It is then possible to identify and select for evaluation those objectives which are common to several projects, and those that are unique to a project. Objectives are typically related to achievement; however, CAM is equally suited to measuring changes in attitudes or perceptions. The pool of objectives is called an objective bank, and a computer program is available to handle the large amount of data involved.

Test items. The second step toward the use of CAM is the construction of test items. Every item is tied specifically to a single objective, and multiple items are constructed for each objective. All items, keyed by objectives, may be stored in a computerized item bank, ready for sampling or available for revision.

Construction of test forms. The number of test forms, or monitors, must at least equal the number of test administrations planned. Tests are made parallel in content by using the technique of stratified random sampling. Forms are also randomly comparable in difficulty. If an item analysis can be run (perhaps on a pretest or an earlier version of the course) for indices of difficulty and discrimination, the forms may be made more exactly comparable in difficulty.

Monitors are intended to be short tests, perhaps ten to thirty items. Whether or not a single form covers all objectives for a course is a function of the proportion of objectives to items-per-form. It may be necessary to randomly sample (without replacement) the objectives, before doing the same on the test items for each selected objective. This technique of sampling must insure that, across forms, all objectives are equally represented. The same consideration holds when items-per-form exceed the number of objectives, in this case, some objectives may be represented by more than one item on some forms.

Student test groups. Students are divided into test groups in order to use all test forms at each administration. Test groups are best constructed using random sampling of strata of students based on ability or prior achievement in the subject. This assures that each group has a range of students which gives representativeness to the data for each test form.

It is most desirable, for several reasons, to include every student in every test administration, and when set up this way, CAM has been found to be a satisfactory substitute for usual classroom testing. However, it is possible to use only a sample of the student population, especially if the number involved in a project approaches one thousand or more. Many different sampling designs are possible. Using the total student population in one test group is the design for the conventional project evaluation. Unequal-sized test groups may sometimes be an administrative necessity.

Test administrations. Test administrations may coincide with the completion of instructional units, or they may be set at regular intervals throughout the course. The latter has advantages in terms of ease of administration, and comparability of results from similar courses taught at different schools.

Appended package tests. It is possible to add a section to any monitor, and have the results incorporated with the rest of the CAM data. This feature

lends flexibility in that, should a specific diagnostic test seem desirable at any point, the data can easily be assimilated.

Data analysis and reporting. Output from the computer programs is as follows:

For individual students

After each administration:

1) total score on that and all previous administrations.

2) a graphic presentation of the above.

3) a right-wrong indication for each item on the monitor, coded by the objective represented.

At the end of the course:

4) average scores, across all monitors taken, on items categorized by use into three groups--pretest, immediate post-instruction and retention of varying lengths of time.

For whole group or subgroups (e.g., one classroom; highest and lowest quartiles)

After each administration:

1) percent answered correctly out of all items across all monitors, for each objective.

Periodically, as desired (e.g., every 3-5 administrations):

2) trend data, or achievement profiles, for total score and for each objective.

At the end of the course

3) same as number 4 under individual students.

4) item analysis (using whole group only), treating each item in three separate ways, by its three functions--pretest, immediate post-instruction, and retention measure.

Data are analyzed, and reports printed, by computer. Data can be collapsed in various ways, to be most useful to students, teachers, project directors, or state evaluators.

Specificity of objectives. Any instruction, no matter how it is to be evaluated, can call for a high degree of specificity of objectives CAN, however,

rigorously prescribes and requires such specificity. It is the base upon which the detailed testing, analysis and feedback of the program rest.

Specificity of objectives allows similar curricula to pool and match their objectives. What is common to all curricula, or to several, is readily observable, and provides a meaningful, detailed comparison. Objectives unique to individual curricula can pinpoint actual differences concretely and precisely.

Test items tied to objectives. Each test item is constructed to measure achievement on a particular objective. Therefore, test data always relate to definite objectives, rather than aggregates of objectives: this allows evaluation procedures to be matched with specific goals of the curricula. In this respect, CAM differs significantly from conventional curriculum project evaluations, where standardized materials are used, which have not been closely tied to the specific objectives for the curriculum.

Modification of curricula. Conventional curriculum project evaluation may provide some criteria upon which to base one kind of decision about an existing project: "drop it" or "continue it". These criteria are global rather than related to specific contributions of the project. Perhaps one of the most valuable characteristics of the comprehensive achievement monitoring model is that it is able to provide information upon which to make specific recommendations for retaining strong components of a project, and modifying weak ones. No project is as effective as possible, as set up at its inception; therefore, a far more pertinent decision about it, now possible with the CAM model, is "drop" or "continue with these modifications."

Data more valid. If there is time on a test for one question for an objective, then estimates of group achievement on that objective will be more valid if a variety of questions is used across the group, rather than the single question typical of both classroom tests and project evaluation. It is important to note that the increased validity and comprehensiveness calls for no sacrifice in the

economy of data collection, since each student need still answer only one question.

Pretest of all objectives. All objectives are pretested before any instruction has been given. First, it is important to know whether students have already acquired information or skills from outside sources, to that the project need not lose students' interest by covering material that they can handle already. Secondly, an index of effectiveness must ultimately be an index related to change in student achievement, attitude or perception. In order to document change, it is necessary to have at least two comparable measurements of the same characteristic, taken at two different times.

There is reason to continue pretesting on objectives to be taught later in the project, because outside learning experiences, or interaction between material taught early in the project and that scheduled to be taught later, may both very reasonably cause changes in performance during the project. This may lead to alterations, either in the sequence of instruction, or the amount of time spent on certain objectives. When the level of achievement rises on an objective not yet taught, it may be closely related to material just taught, in which case, instruction in the later-scheduled unit could be moved up to take pedagogical advantage of the relationship. Another possibility is that, without changing the sequence, certain instructional units might be condensed, and the pace of instruction stepped up. A single pre-course test, will not provide information for making the above decisions.

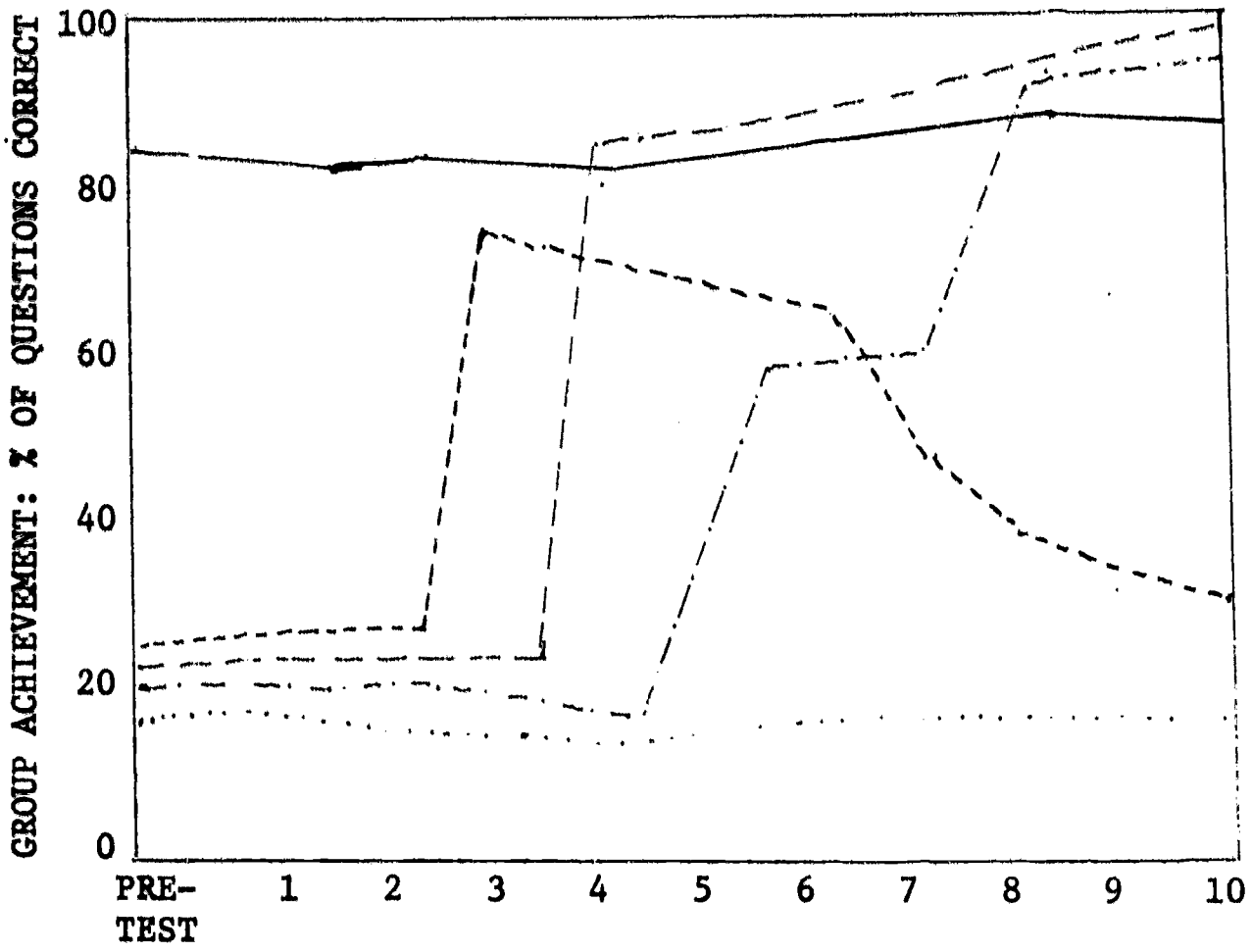
Immediate post-instruction test. The usual classroom test covers only material just taught. CAM estimates of group achievement on just-taught objectives are comparable to those available from classroom testing. The number of students usually involved in projects makes it possible to test each objective with a substantial variety of items, without lengthening any one form of the test.

Continual measure of rentention. Since objectives contineu to be tested after they have been taught, throughout the rest of the course, there is a continual test of retention. Intervals between "teach" and "test" times are of varying length, and can be matched for precise analysis. For example, it would be possible to measure retention spanning appromimately six weeks on all material of a course except what is presented during the last month or so. Therefore, estimates of achievement can be systematically made for each of the instructional units after a specified interval.

Achievement profile. There are comparable data on achievement for every test administration. This makes it possible to plot students' achievement on any given objective (or group of objectives) for the entire course. This plot, called an achievement profile, gives a graphic presentation of the changes in group achievement throughout the course. This achivement profile is a unique characteristic of the information available form the CAI model, and is very useful in describing and reporting results of course and project research.

Figure 1 presents hypothetical achivement profiles for five objectives from a course. Brief comments below the graph give possible interpretations. It is obvious that achievement profiles provide a wealth of information, at whatever point in the course they are drawn. On the pretest in the foregoing example, all objectives except number 2 show achievement at the chance level, or about 20% (five-option multiple-choice items). Several decisions could have been made after test administration one:

1) Objective 1 was not learned--reteach it in some other way; 2) Objective 2 has tested high on both the pretest and test administration 1--it would be safe to skip instruction in this objective. After test administration 5, two other decisions might have been made: 1) Achievement on Objective 3 seems to be slipping--review is needed, preferably soon 2) Objective 8 seems closely related to Objective 5--perhaps it should be taught now instead of later.



.....	Obj. 1:	taught, but students did not learn; with rapid feedback, could be corrected with change in instruction.
——	Obj. 2:	previously known and not taught; without pretest, this looks like student learning.
- - - -	Obj. 3:	taught and learned, but forgotten
— —	Obj. 4:	well taught
—.—.—	Obj. 8:	appears related to objective 5, because achievement increases when 5 is taught.

Figure 1. Hypothetical achievement profiles of group achievement on five objectives.

Continuous data available. Data are available from every test administration. It is possible to look at group achievement on a single objective, groups of objectives, or total content of a course, though this last is generally less useful. Data can be summarized in a variety of ways, through the use of selected computer programs now available. Desired data are always available within a few days for decision-making; it is not necessary to wait weeks or months for meaningful analyses. Many evaluation systems are not able to analyze and report results with sufficient speed and organization to make the information most useful to its recipients. Analyses can be tailor-made for project directors or state evaluators.

One economic advantage of periodic feedback is that a project need not continue to its end to discover, after all funds are spent, that the goals of the project have not been accomplished. Modifications can be made in the program if student performance does not move in the expected direction.

Description of Usual Classroom Testing

The usual classroom testing situation includes the following sequence of events: first, a set of objectives if specified for a limited instructional period, usually from one to four weeks; second, an instructional treatment is devised and administered to the students; and lastly, a test at the close of the instructional period is administered to measure the extent to which the objectives taught during that period have been achieved.

Students' achievement on material taught during instructional period one is tested at test administration one. Achievement for period two is tested at administration two, and so on, throughout the course.

There is usually a "final test" administered at the end of the course, for which there may be varying amounts of review offered. Sometimes major tests are administered at other times during the course e.g. just before report cards are issued.

Flexible weighting. There is great flexibility in the relative emphasis accorded various objectives during the year. Decisions may be made at any time; content may be added, dropped or modified. The testing is tailored to the content as the course progresses.

Individual student testing. Usual classroom testing can yield diagnostic data on individual student achievement, on the few specific objectives which have been taught.

Tests related to objectives. Usual classroom testing may meet the criterion of close relationship between objectives and test items, when the school program is defined in behavioral objectives, and the teacher makes some effort to relate the items directly to the objectives.

No pretesting. There is usually no pretest information on students' prior achievement on any objective. Teachers usually assume that student achievement is due solely to the instruction given them in class. Furthermore, they do not know whether learning one objective has affected understanding of another objective. Students may also have experiences in other courses, or outside of school, either before or during a course, which contribute to their understanding of various objectives, whether or not they have been taught yet.

No test of retention. There is no information on students' retention of objectives which have been taught earlier in the school year, except in the event of some sort of major test. At that test administration, the interval between time of instruction, and time of test-of-retention, is different for every objective taught. The interval may span almost a full school year, or be only a week or two. There is seldom any data attached to such test results about the date of instruction on a given objective.

No comparison of student achievement over time. It is very difficult to compare students' achievement from one point in time to another, because at each test administration, an entirely different test is used; there is seldom any overlap in content, and the overall difficulty can vary enormously from one test to another. The only possible comparison of achievement from one time to another must use a student's rank order in his class. This still leaves no way to examine changes in a total class's achievement over time.

Description of Curriculum Project Evaluation

A frequently used strategy for evaluating curriculum projects is to administer an extensive achievement test at the conclusion of the project. This may consist of a test, or battery of tests, sometimes composed specifically for the project, but usually prepared and distributed commercially, e.g. standardized achievement tests.

There is sometimes a pretest administered before the start of the project, which is either the same as the posttest, or an alternate form of it, but presumes to measure the same objectives.

A single posttest or a pretest-posttest costs less than a more effective and complete evaluation system such as CAM. There is a minimum of clerical and administrative work needed in actually giving the test, and if a commercially available test is used, it may simply be purchased; no staff or time is needed to develop a test tailored to the objectives of the curriculum. What little analysis on results can be done, is relatively easily accomplished.

Deficient immediate post-instruction testing. In terms of immediate post-instruction achievement, the usual curriculum project evaluation measures only the objectives taught at the very end of the project in a way similar to usual classroom testing (i.e., immediately following the instructional treatment). This means

that project directors do not have information on the direct effect of instruction immediately after students have been exposed to it.

Tests of retention. The interval between the teaching of an objective, and the end-of-course test, varies for each objective. Such intervals range from a week or two, to a full school year. Therefore, an estimate of achievement based only on a posttest is an aggregate of immediate post-instruction achievement, short-term retention, and long-term retention. This composite score may be made up of several subscores, but such subscores still do not indicate much about the time interval since instruction.

No comparison of scores. There is no need to discuss comparability of scores from one time to another if the testing is done at only one point in time. Pretest-posttest problems are discussed below under sample attrition.

Test items not specific to objective. In posttests which are designed to cover an entire course at only one administration, there is great variation in the specificity with which test items have been matched to the objectives of the course. This problem is especially apparent when standardized achievement tests are used, where general subscores are roughly matched with the stated objectives of the project. When only standardized tests and materials are used in a post-project evaluation, there is a definite lack of systematic information about the achievement on specific objectives in the program.

Inappropriate weighting. In giving one large posttest, especially a standardized test, the problem of weighting of objectives presents itself. A variety of objectives could be poorly measured while other objectives are heavily emphasized. It is likely that the intended pattern of emphasis in the course will not be reflected in the evaluation instrument.

Test not comprehensive. Not only will there be too little emphasis on certain objectives, but it is possible that some objectives will not be measured at all.

Lack of comprehensiveness in an evaluation technique is a serious shortcoming.

Problems of sample attrition. All of the above weaknesses in the usual curriculum project evaluation design are relatively unimportant when compared with the most serious problem of all: the turnover of students. Those students who were pretested before the program, and received the early segments of instruction, are simply not there at the time of the posttest. Effectively, this reduces the hard data to a posttest on students still enrolled in the project during the final week, even if a pretest were administered. Therefore, the results may represent very little more than immediate post-instruction testing on the objectives taught just before the posttest. Pretest information, if it has been gathered, relates only to the incoming abilities of a sample of students roughly similar to that available for posttesting. The assumption is made that students coming into the project are similar to those leaving it, but the data cannot be used statistically in analyzing changes in student achievement, since change should only be measured for individual students or identical groups of students.

Comparison of the Pattern and Quality of Information of the Models

The amount and quality of information available from the three models of evaluation described above will serve to summarize the characteristics of each.

Comparison of information. CAM yields more information than either the usual classroom testing or conventional curriculum project evaluation. The pattern of data resulting from each model may be fitted into a matrix, in which the rows indicate all the objectives or instructional units of the course, and the columns represent the possible test administrations during the entire project. A cell of the matrix which is filled in, represents an estimate of achievement for that objective or unit, at that test administration.

The usual classroom testing pattern is illustrated in Table 1. The diagonal line of X's at the last administration indicates a final test presumably covering

all the units of the course.

Table 2 illustrates graphically the lack of information available from the usual pretest--posttest curriculum project evaluation. This illustration makes the assumption, not necessarily well-founded, that a single test does in fact provide information about every instructional unit.

It is readily apparent in Table 3 that CAM makes available data on group achievement for all of the objectives specified for a course, at each time of testing. This comprehensiveness of the data provides the necessary information for the variety of purposes discussed earlier in this section. It is easy to see how CAM contrasts with the other models of testing, where information is generally available either on a few of the objectives, or as a composite score for all objectives, at a single time.

Comparison of quality. Table 4 displays seven types of information, and estimates their quality as provided by each of the three models.

Conventional curriculum project evaluation is fair to poor on all of the dimensions described. These shortcomings are inherent in the use of single test long enough to provide detailed information about student performance on a large number of objectives is fatiguing and therefore less valid than short tests. One long test excludes systematic pretest, immediate post-instruction, and detailed retention information. Attrition takes a heavy toll of a pretest sample. Feedback is limited to a post mortem on the project's strengths and weaknesses.

Usual classroom testing provides for the measurement of performance on specific objectives on an immediate post-instruction basis. By repeated testing, the effects of attrition may be minimized. If usual classroom testing data were collected across similar projects after similar objectives had been taught, extensive information would be available for comparing projects. However, an accurate comparison of projects must also include pretest and retention information. The former is used to adjust for incoming aptitude and achievement

TABLE 1

Usual Classroom Testing: Estimates of Achievements Available for a Group of Students by Unit and Test Administration

	Time					
Unit	1	2	3	4	...	T
1	X					X
2		X				X
3			X			X
4				X		X
.					X	X
.						
U						X

TABLE 2

Pretest-Posttest Curriculum Project Evaluation: Estimates of Achievement Available for a Group of Students by Unit and Test Administration.

	Time					
Unit	1	2	3	4	...	T
1	X					X
2	X					X
3	X					X
4	X					X
.	X					X
.						
.						
U	X					X

TABLE 3

Comprehensive Achievement Monitoring Evaluation
Estimates of Achievement Available for a Group
of Students by Unit and Test Administration

Unit	Time 1	2	3	4	...	T
1	C	C	C	C	C	C
2	C	C	C	C	C	C
3	C	C	C	C	C	C
4	C	C	C	C	C	C
. . .	C	C	C	C	C	C
U	C	C	C	C	C	C

TABLE 4

Quality of Information

Available from Three Evaluation Models

Information	Model		
	Usual Classroom Testing	Usual Project Evaluation	Comprehensive Achievement Monitoring
Evaluation specific to objectives	***	**	***
Pretest of objectives	*	*	***
Immediate post-instruction testing	***	**	***
Evaluation of retention of objectives	*	**	***
Comparability across time	*	*	***
Achievement profiles			***
Continuous feedback	**	*	***
Immunity to sample attrition	***	*	***

NOTE: Quality of information rated as excellent(****), good(***), fair (**), poor (*), and not available (blank).

differences in students, and the latter for long-term retention, or payoff of the project. Neither of these is specifically available from classroom testing. Feedback occurs frequently during the project, but provides information about only one instructional unit at a time.

Comprehensive Achievement Monitoring provides information for evaluation comparable, or superior, to the other evaluation models. Its superiority lies in the areas of particular importance to project evaluation: systematic pretests and measures of retention of objectives. Feedback can be provided continuously and comprehensively so that the projects can be critiqued and adjustments made before their end.

Evaluative Issues

Comparability. Similarity of instruments and comprehensiveness of the data generated are necessary to obtain indices of effectiveness on a state-wide basis or within a school district. It would be difficult to observe change in academic achievement if the measures taken at one time, or in one school, were not directly comparable with measures taken at another time or in another school. Further, it is of crucial importance in evaluating educational programs that there be a systematic, on-going, objective accummulation of information about the achievement of all behavioral objectives. Both of these concerns are provided for within the structure of CAM.

Another issue in sharpening the evaluative process and improving the quality of instructional treatment and curriculum design, is that of clearly specifying behavioral objectives and performance criteria. It is inherent in the CAM design that courses be thoughtfully and systematically planned, without, however, destroying opportunities for creative and ad hoc improvisation.

Flexibility. Comparability does not necessitate a single standardized evaluative instrument as CAM has shown. Wide diversity in course structure must be accommodated. What comparability demands is that the objectives of different

programs for the same subject be carefully specified and tested. For it is impossible to compare course achievement levels from school to school, or even from class to class, if the evaluator is ignorant of the dimensions in which they differ. However, carefully specified courses can be compared on their common components by the CAM technique.