14. COURSE IMPROVEMENT THROUGH EVALUATION

LEE J. CRONBACH

The national interest in improving education has generated several highly important projects attempting to improve curricula, particularly at the secondary-school level. In conferences of directors of course content improvement programs sponsored by the National Science Foundation, questions about evaluation are frequently raised.[1] Those who inquire about evaluation have various motives, ranging from sheer scientific curiosity about classroom events to a desire to assure a sponsor that money has been well spent. While the curriculum developers sincerely wish to use the skills of evaluation specialists, I am not certain that they have a clear picture of what evaluation can do and should try to do. And, on the other hand, I am becoming convinced that some techniques and habits of thought of the evaluation specialist are ill-suited to current curriculum studies. To serve these studies, what philosophy and methods of evaluation are required? And, particularly, how must we depart from the familiar doctrines and rituals of the testing game?

DECISIONS SERVED BY EVALUATION

To draw attention to its full range of functions, we may define evaluation broadly as the *collection and use of information to make decisions about an educational program.* This program may be a set of instructional materials distributed nationally, the instruc-

From *Teachers College Record,* 64 (1963), 672–83. Copyright 1963, Teachers College, Columbia University, New York. Reprinted with permission of the author and publisher using an edited version found in R. W. Heath, *New Curricula.* Harper & Row, 1964, at Professor Cronbach's request.

tional activities of a single school, or the educational experiences of a single pupil. Many types of decision are to be made, and many varieties of information are useful. It becomes immediately apparent that evaluation is a diversified activity and that no one set of principles will suffice for all situations. But measurement specialists have so concentrated upon one process—the preparation of pencil-and-paper achievement tests for assigning scores to individual pupils—that the principles pertinent to that process have somehow become enshrined as *the* principles of evaluation. "Tests," we are told, "should fit the content of the curriculum." Also, "only those evaluation procedures should be used that yield reliable scores." These and other hallowed principles are not entirely appropriate to evaluation for course improvement. Before proceeding to support this contention, I wish to distinguish among purposes of evaluation and relate them to historical developments in testing and curriculum making.

We may separate three types of decisions for which evaluation is used:

1. Course improvement: deciding what instructional materials and methods are satisfactory and where change is needed.
2. Decisions about individuals: identifying the needs of the pupil for the sake of planning his instruction, judging pupil merit for purposes of selection and grouping, acquainting the pupil with his own progress and deficiencies.
3. Administrative regulation: judging how good the school system is, how good individual teachers are, etc.

Course improvement is set apart by its broad temporal and geographical reference; it involves the modification of recurrently used materials and methods. Developing a standard exercise to overcome a misunderstanding would be course improvement, but deciding whether a certain pupil should work through that exercise would be an individual decision. Administrative regulation likewise is local in effect, whereas an improvement in a course is likely to be pertinent wherever the course is offered.

It was for the sake of course improvement that systematic evaluation was first introduced. When that famous muckraker Joseph Rice gave the same spelling test in a number of American schools and so gave the first impetus to the educational testing movement, he was interested in evaluating a curriculum. Crusading against the extended spelling drills that then loomed large in the school schedule—"the spelling grind"—Rice collected evidence of their worthlessness so as to provoke curriculum revision. As the testing movement developed, however, it took on a different function.

The greatest expansion of systematic achievement testing occurred in the 1920s. At that time, the content of any course was taken pretty much as established and beyond criticism, save for small shifts of topical emphasis. At the administrator's direction, standard tests covering this curriculum were given to assess the efficiency of the teacher or the school system. Such administrative testing fell into disfavor when used injudiciously and heavy-handedly in the 1920s and 1930s. Administrators and accrediting agencies fell back upon descriptive features of the school program in

judging adequacy. Instead of collecting direct evidence of educational impact, they judged schools in terms of size of budget, student-staff ratio, square feet of laboratory space, and the number of advanced credits accumulated by the teacher. This tide, it appears, is about to turn. On many university campuses, administrators wanting to know more about their product are installing "operations research offices." Testing directed toward quality control seems likely to increase in the lower schools as well, as is most forcefully indicated by the statewide testing just ordered by the California legislature.

After 1930 or thereabouts, tests were given almost exclusively for judgments about individuals: to select students for advanced training, to assign marks within a class, and to diagnose individual competences and deficiencies. For any such decisions, one wants precise and valid comparisons of one individual with other individuals or with a standard. Much of test theory and test technology has been concerned with making measurements precise. Important though precision is for most decisions about individuals, I shall argue that in evaluating courses we need not struggle to obtain precise scores for individuals.

While measurers have been well content with the devices used to make scores precise, they have been less complacent about validity. Prior to 1935, the pupil was examined mostly on factual knowledge and mastery of fundamental skills. Tyler's research and writings of that period developed awareness that higher mental processes are not evoked by simple factual tests and that instruction that promotes factual knowledge may not promote—indeed, may interfere with—other more important educational outcomes. Tyler, Lindquist, and their students demonstrated that tests can be designed to measure general educational outcomes, such as ability to comprehend scientific method. Whereas a student can prepare for a factual test only through a course of study that includes the facts tested, many different courses of study may promote the same *general* understandings and attitudes. In evaluating today's new curricula, it will clearly be important to appraise the student's general educational growth, which curriculum developers say is more important than mastery of the specific lessons presented. Note, for example, that the Biological Sciences Curriculum Study offers three courses with substantially different "subject matter" as alternative routes to much the same educational ends.

Although some instruments capable of measuring general outcomes were prepared during the 1930s, they were never very widely employed. The prevailing philosophy of the curriculum, particularly among progressives, called for developing a program to fit local requirements, capitalizing on the capacities and experiences of local pupils. The faith of the 1920s in a "standard" curriculum was replaced by a faith that the best learning experience would result from teacher-pupil planning in each classroom. Since each teacher or each class could choose different content and even different objectives, this philosophy left little place for standard testing.

Many evaluation specialists came to see test development as a strategy for training the teacher in service, so that the process of test making came to be valued more than the test—or the test data—that resulted. The following remarks by Bloom (1961) are representative of a whole school of thought:[2]

> The criterion for determining the quality of a school and its educational functions would be the extent to which it achieves the objectives it has set for itself. . . . (Our experiences suggest that unless the school has translated the objectives into specific and operational definitions, little is likely to be done about the objectives. They remain pious hopes and platitudes.) . . . Participation of the teaching staff in selecting as well as constructing evaluation instruments has resulted in improved instruments on one hand, and, on the other hand, it has resulted in clarifying the objectives of instruction and in making them real and meaningful to teachers. . . . When teachers have actively participated in defining objectives and in selecting or constructing evaluation instruments, they return to the learning problems with great vigor and remarkable creativity. . . . Teachers who have become committed to a set of educational objectives which they thoroughly understand respond by developing a variety of learning experiences which are as diverse and as complex as the situation requires.

Thus "evaluation" becomes a local, and beneficial, teacher-training activity. The benefit is attributed to thinking about the data to collect. Little is said about the actual use of test results; one has the impression that when test-making ends, the test itself is forgotten. Certainly there is little enthusiasm for refining tests so that they can be used in other schools, for to do so would be to rob those teachers of the benefits of working out their own objectives and instruments.

Bloom and Tyler describe both curriculum making and evaluation as integral parts of classroom instruction, which is necessarily decentralized. This outlook is far from that of course improvement. The current national curriculum studies assume that curriculum making can be centralized. They prepare materials to be used in much the same way by teachers everywhere. It is assumed that having experts draft materials and revising these after tryout produces better instructional activities than the local teacher would be likely to devise. In this context, it seems wholly appropriate to have most tests prepared by a central staff and to have results returned to that staff to guide further course improvement.

When evaluation is carried out in the service of course improvement, the chief aim is to ascertain what effects the course has—that is, what changes it produces in pupils. This is not to inquire merely whether the course is effective or ineffective. Outcomes of instruction are multidimensional, and a satisfactory investigation will map out the effects of the course along these dimensions separately. To agglomerate many types of post-course performance into a single score is a mistake, since failure to achieve one objective is masked by access in another direction. Moreover, since a composite score embodies (and usually conceals) judgments about the importance of the various outcomes, only a report that treats the outcomes separately can be useful to educators who have different value hierarchies.

The greatest service evaluation can perform is to identify aspects of the course where revision is desirable. Those responsible for developing a course would like to present evidence that their course is effective. They are intrigued by the idea of having an "independent testing agency" render a judgment on their product, but to call in the evaluator only upon the completion of course development, to confirm what has been done, is to offer him a menial role and make meager use of his services. To

be influential in course improvement, evidence must become available midway in curriculum development, not in the home stretch when the developer is naturally reluctant to tear open a supposedly finished body of materials and techniques. Evaluation, used to improve the course while it is still fluid, contributes more to improvement of education than evaluation used to appraise a product already placed on the market.

Insofar as possible, evaluation should be used to understand how the course produces its effects and what parameters influence its effectiveness. It is important to learn, for example, that the outcome of programed instruction depends very much upon the attitude of the teacher; indeed, this may be more important than to learn that on the average such instruction produces slightly better or worse results than conventional instruction.

Hopefully, evaluation studies will go beyond reporting on this or that course and help us to understand educational learning. Such insight will in the end contribute to the development of all courses rather than just of the course under test. In certain of the new curricula, there are data to suggest that aptitude measures correlate much less with end-of-course achievement than they do with achievement on early units (Ferris, 1962). This finding is not well-confirmed, but is highly significant if true. If it is true for the new curricula and only for them it has one implication; if the same effect appears in traditional courses, it means something else. Either way, it provides food-for-thought for teachers, counselors, and theorists. Evaluation studies should generate knowledge about the nature of the abilities that constitute educational goals. Twenty years after the Eight-Year Study of the Progressive Education Association, its testing techniques are in good repute, but we still know very little about what these instruments measure. Consider "Applications of Principles in Science." Is this in any sense a unitary ability? Or has the able student only mastered certain principles one-by-one? Is the ability demonstrated on a test of this sort more prognostic of any later achievement than is factual knowledge? Such questions ought to receive substantial attention, though to the makers of any one course they are of only peripheral interest.

The aim of comparing one course with another should not dominate plans for evaluation. To be sure, decisionmakers have to choose between courses, and any evaluation report will be interpreted in part comparatively. But formally designed experiments pitting one course against another are rarely definitive enough to justify their cost. Differences between average test scores resulting from different courses are usually small, relative to the wide differences among and within classes taking the same course. At best, an experiment never does more than compare the present version of one course with the present version of another. A major effort to bring the losing contender nearer to perfection would be very likely to reverse the verdict of the experiment.

Any failure to equate the classes taking the competing courses will jeopardize the interpretation of an experiment, and such failures are almost inevitable. In testing a drug, we know that valid results cannot be obtained without a double-blind control, in which the doses for half the subjects are inert placebos; the placebo and the drug look alike, so that neither doctor nor patient knows who is receiving medication.

Without this control, the results are useless even when the state of the patient is checked by completely objective indices. In an educational experiment, it is difficult to keep pupils unaware that they are an experimental group, and it is quite impossible to neutralize the biases of the teacher as those of the doctor are neutralized in the double-blind design. It is thus never certain whether any observed advantage is attributable to the educational innovation, as such, or to the greater energy that teachers and students put forth when a method is fresh and experimental. Some have contended that any course, even the most excellent, loses much of its potency as soon as success enthrones it as the traditional method.[3]

Since group comparisons give equivocal results, I believe that a formal study should be designed primarily to determine the post-course performance of a well-described group, with respect to many important objectives and side effects. Ours is a problem like that of the engineer examining a new automobile. He can set himself the task of defining its performance characteristics and its dependability. It would be merely distracting to put his question in the form: "Is this car better or worse than the competing brand?" Moreover, in an experiment where the treatments compared differ in a dozen respects, no understanding is gained from the fact that the experiment shows a numerical advantage in favor of the new course. No one knows which of the ingredients is responsible for the advantage. More analytic experiments are much more useful than field trials applying markedly dissimilar treatments to different groups. Small-scale, well-controlled studies can profitably be used to compare alternative versions of the same course; in such a study the differences between treatments are few enough and well-enough defined that the results have explanatory value.

The three purposes—course improvement, decisions about individuals, and administrative regulation—call for measurement procedures having somewhat different qualities. When a test will be used to make an administrative judgment on the individual teacher, it is necessary to measure thoroughly and with conspicuous fairness; such testing, if it is to cover more than one outcome, becomes extremely time-consuming. In judging a course, however, one can make satisfactory interpretations from data collected on a sampling basis, with no pretense of measuring thoroughly the accomplishments of any one class. A similar point is to be made about testing for decisions about individuals. A test of individuals must be conspicuously fair and extensive enough to provide a dependable score for each person. But if the performance will not influence the fate of the individual, we can ask him to perform tasks for which the course has not directly prepared him, and we can use techniques that would be prohibitively expensive if applied in a manner thorough enough to measure each person reliably.

METHODS OF EVALUATION

Range of Methods

Evaluation is too often visualized as the administration of a formal test, an hour or so in duration, at the close of a course. But there are many other methods for exam-

ining pupil performance, and pupil attainment is not the only basis for appraising a course.

It is quite appropriate to ask scholars whether the statements made in the course are consistent with the best contemporary knowledge. This is a sound, even a necessary, procedure. One might go on to evaluate the pedagogy of the new course by soliciting opinions, but here there is considerable hazard. If the opinions are based on some preconception about teaching method, the findings will be control versial and very probably misleading. There are no theories of pedagogy so well established that one can say, without tryout, what will prove educative.

One can accept the need for a pragmatic test of the curriculum and still employ opinions as a source of evidence. During the tryout stages of curriculum making one relies heavily on the teachers' reports of pupil accomplishment—"Here they had trouble"; "This they found dull"; "Here they needed only half as many exercises as were provided"; etc. This is behavior observation, even though unsystematic, and it is of great value. The reason for shifting to systematic observation is that this is more impartial, more public, and sometimes more penetrating. While I bow to the historian or mathematician as a judge of the technical soundness of course content, I do not agree that the experienced history or mathematics teacher who tries out a course gives the best possible judgment of its effectiveness. Scholars have too often deluded themselves about their effectiveness as teachers—in particular, they have too often accepted parroting of words as evidence of insight—for their unaided judgment to be trusted. Systematic observation is costly and introduces some delay between the moment of teaching and the feedback of results. Hence, systematic observation will never be the curriculum developer's sole source of evidence. Systematic data collection becomes profitable in the intermediate stages of curriculum development, after the more obvious bugs in early drafts have been dealt with.

The approaches to evaluation include process studies, proficiency measures, attitude measures, and follow-up studies. A process study is concerned with events taking place in the classroom, proficiency and attitude measures with changes observed in pupils, and follow-up studies with the later careers of those who participated in the course.

The follow-up study comes closest to observing ultimate educational contributions, but the completion of such a study is so far removed in time from the initial instruction that it is of minor value in improving the course or explaining its effects. The follow-up study differs strikingly from the other types of evaluation study in one respect. I have already expressed the view that evaluation should be primarily concerned with the effects of the course under study rather than with comparisons of courses. That is to say, I would emphasize departures of attained results from the ideal, differences in apparent effectiveness of different parts of the course, and differences from item to item. All these suggest places where the course could be strengthened; but this view cannot be applied to the follow-up study, which appraises effects of the course as a whole and which has very little meaning unless outcomes can be compared with some sort of base rate. Suppose we find that 65 percent of

the boys graduating from an experimental curriculum enroll in scientific and technical majors in college. We cannot judge whether this is a high or low figure save by comparing it with the rate among boys who have not had this course. In a follow-up study, it is necessary to obtain data on a control group equated at least crudely to the experimental cases on the obvious demographic variables.

Despite the fact that such groups are hard to equate and that follow-up data do not tell much about how to improve the course, such studies should have a place in research on the new curricula, whose national samples provide unusual opportunity for follow-up that can shed light on important questions. One obvious type of follow-up study traces the student's success in a college course founded upon the high school course. One may examine the student's grades or ask him what topics in the college course he found himself poorly prepared for. It is hoped that some of the new science and mathematics courses will arouse greater interest than usual among girls; whether this hope is well-founded can be checked by finding out what majors and what electives these ex-students pursue in college. Career choices likewise merit attention. Some proponents of the new curricula would like to see a greater flow of talent into basic science as distinct from technology, while others would regard this as potentially disastrous; but no one would regard facts about this flow as lacking significance.

Attitudes are prominent among the outcomes that course developers are concerned with. Attitudes are meanings or beliefs, not mere expressions of approval or disapproval. One's attitude toward science includes ideas about the matters on which a scientist can be an authority—about the benefits to be obtained from moon shots and studies of monkey mothers, and about depletion of natural resources. Equally important is the match between self-concept and concept of the field: what roles does science offer a person like me? Would I want to marry a scientist? and so on. Each learning activity also contributes to attitudes that reach far beyond any one subject, such as the pupil's sense of his own competence and desire to learn.

Attitudes can be measured in many ways; the choices revealed in follow-up studies, for example, are pertinent evidence. But measurement usually takes the form of direct or indirect questioning. Interviews, questionnaires, and the like are quite valuable when not trusted blindly. Certainly, we should take seriously any *un*desirable opinion expressed by a substantial proportion of graduates of a course (e.g., the belief that the scientist speaks with peculiar authority on political and ethical questions, or the belief that mathematics is a finished subject rather than a field for current investigation).

Attitude questionnaires have been much criticized because they are subject to distortion, especially where the student hopes to gain by being less than frank. Particularly if the questions are asked in a context far removed from the experimental course, the returns are likely to be trustworthy. Thus, a general questionnaire administered through homerooms (or required English courses) may include questions about liking for various subjects and activities; these same questions administered by the mathematics teacher would give much less trustworthy data on attitudes toward mathematics. While students may give reports more favorable than their true beliefs,

this distortion is not likely to be greater one year than another or greater among students who take an experimental course than among those who do not. In group averages, many distortions balance out. But questionnaires insufficiently valid for individual testing can be used in evaluating curricula, both because the student has little motive to distort and because the evaluator is comparing averages rather than individuals.

For measuring proficiency, techniques are likewise varied. Standardized tests are useful, but for course evaluation it makes sense to assign *different* questions to different students. Giving each student in a population of 500 the same test of 50 questions will provide far less information to the course developer than drawing for each student 50 questions from a pool of, say, 700. The latter plan determines the mean success of about 75 representative students on every one of the 700 items; the former reports on only 50 items (See Lord, 1962). Essay tests and open-ended questions, generally too expensive to use for routine evaluation, can profitably be employed to appraise certain abilities. One can go further and observe individuals or groups as they attack a research problem in the laboratory or work through some other complex problem. Since it is necessary to test only a representative sample of pupils, costs are not as serious a consideration as in routine testing. Additional aspects of proficiency testing will be considered below.

Process measures have especial value in showing how a course can be improved, because they examine what happens during instruction. In the development of programed instructional materials, for example, records are collected showing how many pupils miss each item presented; any piling up of errors implies a need for better explanation or a more gradual approach to a difficult topic. Immediately after showing a teaching film, one can interview students, perhaps asking them to describe a still photograph taken from the film. Misleading presentations, ideas given insufficient emphasis, and matters left unclear will be identified by such methods. Similar interviews can disclose what pupils take away from a laboratory activity or a discussion. A process study might turn attention to what the teacher does in the classroom. In those curricula that allow choice of topics, for example, it is worthwhile to find out which topics are chosen and how much time is allotted to each. A log of class activities (preferably recorded by a pupil rather than the teacher) will show which of the techniques suggested in a summer institute are actually adopted, and which form part of the new course only in the developer's fantasies.

Measurement of Proficiency

I have indicated that I consider item data to be more important than test scores. The total score may give confidence in a curriculum or give rise to discouragement, but it tells very little about how to produce further improvement. And, as Ferris (1962) has noted, such scores are quite likely to be mis- or overinterpreted. The score on a single item or on a problem that demands several responses in succession is more likely than the test score to suggest how to alter the presentation. When we accept item scores as useful, we need no longer think of evaluation as a one-shot, end-of-year operation. Proficiency can be measured at any moment,

with particular interest attaching to those items most related to the recent lessons. Other items calling for general abilities can profitably be administered repeatedly during the course (perhaps to different random samples of pupils) so that we can begin to learn when and from what experiences change in these abilities comes.

In course evaluation, we need not be much concerned about making measuring instruments fit the curriculum. However startling this declaration may seem and however contrary to the principles of evaluation for other purposes, this must be our position if we want to know what changes a course produces in the pupil. An ideal evaluation would include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which this curriculum directs substantial attention. If you wish only to know how well a curriculum is achieving *its* objectives, you fit the test to the curriculum; but if you wish to know how well the curriculum is serving the national interest, you measure all outcomes that might be worth striving for. One of the new mathematics courses might disavow any attempt to teach numerical trigonometry and, indeed might discard nearly all computational work. It is still perfectly reasonable to as how well graduates of the course can compute and can solve right triangles. Even the course developers went so far as to contend that computational skill is not proper objective of secondary instruction, they will encounter educators and laymen who do not share their view. If it can be shown that students who come through the new course are fairly proficient in computation despite the lack of direct teaching, the doubters will be reassured. If not, the evidence makes clear how much is being sacrificed. Similarly, when the biologists offer alternative courses emphasizing microbiology and ecology, it is fair to ask how well the graduate of one course can understand issues treated in the other. Ideal evaluation in mathematics will collect evidence on all the abilities toward which a mathemation course might reasonably aim, likewise in biology, English, or any other subject.

Ferris states that the ACS Chemistry Test, however well constructed, inadequate for evaluating the new CBA and CHEM programs, because it does not cover their objectives. One can agree with this without regarding the ACS test inappropriate to use with these courses. It is important that this test not stand aline as the sole evaluation device. It will tell us something worth knowing, namely, just how much "conventional" knowledge the new curriculum does or does not provide. The curriculum developers deliberately planned to sacrifice some of the conventional attainments and have nothing to fear from this measurment, if it competently interpreted (particularly if data are examined item-by-item).

The demand that tests be closely matched to the aims of a course reflect awareness that examinations of the usual sort "determine what is taught." questions are known in advance, students give more attention to learning the answers than to learning other aspects of the course. This is not necessarily detrimental. Wherever it is critically important to master certain content, the knowledge that it will be tested produces a desirable concentration of effort. On the other hand, learning the answer to a set question is by no means the same acquiring understanding of what-

ever topic that question represents. There is therefore, a possible advantage in using "secure" tests for course evaluation. Security is achieved only at a price: one must prepare new tests each year a cannot make before-and-after comparisons with the same items. One would how that the use of different items with different students and the fact that there is low incentive to coach when no judgment is to be passed on the pupils and the teacher would make security a less critical problem.

The distinction between factual tests and tests of higher mental processes, is elaborated for example in the *Taxonomy of Educational Objectives,* is of some value in planning tests, although classifying items as measures of knowledge, application, original problem solving, etc., is difficult and often impossible. Whether a given response represents rote recall of reasoning depends upon how the pupil has been taught, not solely upon the question asked. One might, for example, describe a biological environment and ask for predictions regarding the effect of a certain intervention. Students who had never dealt with ecological data would succeed or fail according to their general ability to reason about complex events; those who had studied ecological biology would be more likely to succeed, reasoning from specific principles; and those who had lived in such an ecology or read about it might answer successfully on the basis of memory. We rarely, therefore, will want to test whether a student *knows* or *does not know* certain material. Knowledge is a matter of degree. Two persons may be acquainted with the same facts or principles, but one will be more expert in his understanding, better able to cope with inconsistent data, irrelevant sources of confusion, and apparent exceptions to the principle. To measure intellectual competence is to measure depth, connectedness, and applicability of knowledge.

Too often, test questions are course-specific, stated in such a way that only the person who has been specifically taught to understand what is being asked for can answer the question. Such questions can usually be identified by their use of conventions. Some conventions are commonplace, and we can assume that all the pupils we test will know them. But a biology test that describes a metabolic process with the aid of the symbol presents difficulties for students who can think through the scientific question about equilibrium but are unfamiliar with the symbol. A trigonometry problem that requires use of a trigonometric table is unreasonable, unless we want to test familiarity with the conventional names of functions. The same problem in numerical trigonometry can be cast in a form clear to the average pupil *entering* high school; if necessary, the tables of functions can be presented along with a comprehensible explanation. So stated, the problem becomes course-independent. It is fair to ask whether graduates of the experimental course can solve such problems, not previously encountered, whereas it is pointless to ask whether they can answer questions whose language is strange to them. To be sure, knowledge of a certain terminology is a significant objective of instruction; but, for course evaluation, testing of terminology should very likely be separated from testing of other understandings. To appraise understanding of processes and relations, the fair question is one comprehensible to a pupil who has not taken the course. This is not to say that he should know the answer or the procedure to follow in attaining

the answer, but he should understand what he is being asked. Such course-independent questions can be used as standard instruments to investigate any instructional program.

Pupils who have not studied a topic usually will be less facile than those who have studied it. Graduates of my hypothetical mathematics course will take longer to solve trigonometry problems than will those who have studied trigonometry. But speed and power should not be confused; in intellectual studies, power, is almost always of greatest importance. If the course equips the pupil to deal correctly, even though haltingly, with a topic not studied, we can expect him to develop facility later when that topic comes before him frequently.

The chief objective in many of the new curricula seems to be to develop aptitude for mastering new materials in the field. A biology course cannot cover all valuable biological content, but it may reasonably aspire to equip the pupil to understand descriptions of unfamiliar organisms, to comprehend a new theory and the reasoning behind it, and to plan an experiment to test a new hypothesis. This is transfer of learning. It has been insufficiently recognized that there are two types of transfer. The two types shade into one another, being arranged on a continuum of immediacy of effect; we can label the more immediate pole *applicational transfer,* and speak of slower-acting effects as *gains in aptitude* (Ferguson, 1954).

Nearly all educational research on transfer has tested immediate performance on a partly new task. We teach pupils to solve equations in x and include in the test equations stated in a or z. We teach the principles of ecological balance by referring to forests and, as a transfer test, ask what effect pollution will have on the population of a lake. We describe an experiment not presented in the text and ask the student to discuss possible interpretations and needed controls. Any of these tests can be administered in short time. But the more significant type of transfer may be the increased ability to learn in a particular field. There is very likely a considerable difference between the ability to draw conclusions from a neatly finished experiment and the ability to tease insight out of the disordered and inconsistent observations that come with continuous laboratory work on a problem. The student who masters a good biology course may become better able to comprehend certain types of theory and data, so that he gains more from a subsequent year of study in ethnology; we do not measure this gain by testing his understanding of short passages in ethnology. There has rarely been an appraisal of ability to work through a problem situation or a complex body of knowledge over a period of days or months. Despite the practical difficulties that attend an attempt to measure the effect of a course on a person's subsequent learning, such *learning to learn* is so important that a serious effort should be made to detect such effects and to understand how they may be fostered.

The technique of programed instruction may be adopted to appraise learning ability. One might, for example, test the student's rate of mastery of a self-contained, programed unit on heat or some other topic not studied. If the program is truly self-contained, every student can master it, but the one with greater scientific com-

prehension hopefully will make fewer errors and progress faster. The program might be prepared in several logically complete versions, ranging from one with very small steps to one with minimal internal redundancy, on the hypothesis that the better-educated student could cope with the less redundant program. Moreover, he might prefer its greater elegance.

CONCLUSION

Old habits of thought and long-established techniques are poor guides to the evaluation required for course improvement. Traditionally, educational measurement has been chiefly concerned with producing fair and precise scores for comparing individuals; educational experimentation has been concerned with comparing score averages of competing courses; but course evaluation calls for description of outcomes. This description should be made on the broadest possible scale, even at the sacrifice of superficial fairness and precision.

Course evaluation should ascertain what changes a course produces and should identify aspects of the course that need revision. The outcomes observed should include general outcomes ranging far beyond the content of the curriculum itself: attitudes, careers choices, general understandings and intellectual powers, and aptitude for further learning in the field. Analysis of performance on single items or types of problems is more informative than analysis of composite scores. It is not necessary or desirable to give the same test to all pupils; rather, as many questions as possible should be given, each to a different moderate-sized sample of pupils. Costly techniques, such as interviews and essay tests, can be applied profitably to samples of pupils, whereas testing everyone would be out of the question.

Asking the right questions about educational outcomes can do much to improve educational effectiveness. Even if the right data are collected, evaluation will have contributed too little if it only places a seal of approval on certain courses and casts others into disfavor. Evaluation is a fundamental part of curriculum development, not an appendage. Its job is to collect facts the course developer can and will use to do a better job and facts from which a deeper understanding of the educational process will emerge.

NOTES

1. My comments on these questions and on certain more significant questions that *should* have been raised, have been greatly clarified by the reactions of several of these directors and colleagues in evaluation to a draft of this paper. J. Thomas Hastings and Robert Heath have been especially helpful. What I voice, however, are my personal views, deliberately more provocative than *authoritative.*

2. Elsewhere, Bloom's paper discusses evaluation for the new curricula. Attention may also be drawn to Tyler's highly pertinent paper (1951).

3. The interested reader can find further striking parallels between curriculum studies and drug research (see Modell, 1963).