DOCUMENT RESUME

ED 141 375 TM 006 005

AUTHOR Gage, N. L.

TITLE The Nature of the Evaluation Process. IIEP Seminar Paper 1.

INSTITUTION United Nations Educational, Scientific, and Cultural Organization, Paris (France). International Inst. for Educational Planning.

PUB DATE 75

NOTE 17p.; Paper presented at the International Institute for Educational Planning Seminar on the Evaluation of the Qualitative Aspects of Education (Paris, France, September 30-October 4, 1974); For related documents, see TM CC6 C03-CC7

AVAILABLE FRCM International Institute for Educational Planning, 7-9 Rue Eugene-Delacroix 75016, Paris, France (3.00 francs)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.

DESCRIPTORS Academic Achievement; Educational Programs; *Evaluation Methods; Formative Evaluation; *Guides; *Program Evaluation; Summative Evaluation

ABSTRACT

Seven major steps are outlined in this brief introduction to the process of evaluation: (1) Identification of the cbjectives of the educational enterprise. (2) Specification of the types of student behavior that reflect achievement of the objectives. (3) Construction of situaticns in which the student will be required or expected to demonstrate desired ways of behaving. (4) Determination of the criteria that will be used to assess the value, correctness, or desirability of the behavior elicited. (5) Administration of the measuring instrument--the tests, rating scales, situational performance tests, etc. (6) Scoring the responses elicited by the measuring instruments. (7) Judging the degree to which the scores obtained by the students reflect achievement of the educational activity. The author warns that misunderstandings about the function of evaluation arise from the failure to distinguish between summative and formative evaluation. The former comes at the end of an educational activity or project, while the latter occurs continually during the activity or project, with the aim of providing feedback to ensure that the appropriate aims and objectives are being met. (Author/MV)

IIEP seminar paper: 1

THE NATURE OF THE EVALUATION PROCESS

N.L. Gage

A contribution to the IIEP Seminar
on "The evaluation of the qualitative
aspects of education"
30 September - 4 October 1974

INTERNATIONAL INSTITUTE
FOR EDUCATIONAL PLANNING
(established by Unesco)
7-9, rue Eugène-Delacroix, 75016 Paris

The opinions expressed in these papers are those of the authors and do not necessarily represent the views of the Institute or of Unesco.

CONTENTS

THE NATURE OF THE EVALUATION PROCESS(1)

The evaluation process is centrally involved in the conduct and improvement of education. This paper provides a brief characterisation of that process. This characterisation is intended to explicate the ways in which the evaluation of educational achievement has come to be carried out as a result of contemporary theoretical and practical work.

MAJOR STEPS IN THE EVALUATION PROCESS

In general, the process of evaluation consists of determining the degree and character of the value of something. In education, evaluation refers to the process of determining the degree to which the objectives of an educational activity, or enterprise, have been achieved. The activity may be a large one, in terms of its duration and scope, such as an entire course or even a curriculum for producing persons with certain kinds of professional or general skill. Or the activity may be relatively restricted, such as the work of a single instructor with his group of adults over a period of a week or an hour. In either case, the evaluation process determines the degree to which the 'objectives' -- here used as a synonym of purposes, goals, aims, etc. -- have been achieved.

As it has come to be practised in education, the process of evaluation has usually consisted of the following steps, here listed in the logical, and what is also typically the chronological, order in which these steps are carried out. Different writers would vary in the specific ways in which these steps are clustered or grouped, but the essentials are well established.

Step 1. Identify the objectives of the educational activity or enterprise.

Often, perhaps usually, the objectives are stated in terms of desired changes in the ways in which students can behave (their capabilities) or will typically behave (their habits or tendencies). This way of stating objectives has several advantages over such alternatives as stating objectives in terms of teacher activities or educational facilities. These advantages result from focusing on the ends (student achievement of various capabilities and tendencies) rather than the various means (e.g. teacher activities or educational facilities) by which these ends will be sought. It is a distinction between ultimate and intermediate objectives. Once the intermediate objectives

(1) This paper was originally written for an IIEP report to the World Bank on the evaluation of qualitative aspects of the educational process. Other authors who contributed to this chapter were T.N. Postlethwaite, Bruce Choppin, Arieh Lewy and Douglas Pidgeon.

have been attained and, for example, educational facilities have been constructed and teachers have been trained, evaluation must become concerned with whether their purposes - certain kinds of student achievement, capability, and habits - are being attained.

Thus, in the one education project, it was stated that "Immediate needs of the education system are the expansion of teacher training, medical education, training of skilled craftsmen, and the integration of various types of rural education and training schemes (including) work-oriented education and training for youths and adults... ." This kind of statement of objective should be converted, insofar as possible, into statements of desired changes in students' abilities and tendencies to behave in certain ways, e.g., perform such tasks as are entailed in various kinds of farming and skilled crafts.

Step 2. Specify the types of student behaviour that will be considered to reflect achievement of the objectives. In the cognitive domain, the student behaviour would consist of ability to remember, comprehend, apply, analyze, synthesize, or evaluate ideas, facts, concepts, principles, problems, and the like. These various kinds of abilities should be defined in observable terms. Thus, instead of resting with such terms as 'knowledge, comprehension, critical thinking, or grasping the significance of', one should attempt to use terms that imply observable behaviour, such as'stating, recognizing, distinguishing true statements from false, matching, putting into one's own words, computing, naming, stating relationships between, or listing the consequences of'.

In the affective domain, the student behaviour would consist of various degrees of internalization of, and commitment to, attitudes and values considered desirable. Here again, an effort to maximize observability should be made. Thus, instead of such terms as 'appreciate, have an interest in, respect, etc.', one might use such terms as 'on his own initiative, he seeks out, tries, approaches, speaks favourably of, spends his own money for, pays attention to, etc.'.

In the psychomotor domain, the student behaviour readily takes such observable forms as ways of moving large and small limbs, communicating and expressing non-verbally, using tools and equipment, operating machines, and motor aspects of ways of speaking and writing.

Step 3. Construct situations in which the student will be required or expected to demonstrate the desired ways of behaving. The situations may range from simple and short questions (or 'items') that will elicit a kind of knowledge to more elaborate problems that will call for a higher mental process (e.g., synthesizing or evaluating). The situations may be verbal, numerical, mechanical, spatial, esthetic, social, or whatever - the main criterion being relevance to the objectives. The situations may be contrived entirely or the kind that occur regularly in 'real life'.

In devising such situations, it is often desirable to make surveys of representative samples of the real-life conditions and circumstances under which the desired ways of behaving would be exhibited. Thus, when it comes to certain kinds of farming skills, the objectives should be defined on the basis of surveys of the kinds of soils, crops, terrain, and the like, in which the farming will be done. For various kinds of skilled crafts, surveys should be made of representative samples of the kinds of carpentry, wood, joints, structures, etc., that a carpenter would need to deal with in the work for which he was being trained. Similarly, the kinds of habits of accuracy, neatness, punctuality, and dependability entailed in the successful performance of the work would be specified. All of the specifications should be based, insofar as possible, on carefully designed surveys of representative samples of the real-life situations in which the student or trainee will subsequently work.

For certain kinds of work, various kinds of general knowledge, intellectual capabilities, and habits are required. The definition of observable behaviours that will cover these kinds of general requirements of living and working in a society should be based on a broad knowledge of the social and cultural characteristics of that society. Here we refer to such matters as literacy and numeracy, health habits of various kinds, and other kinds of behaviour necessary to fitting in and contributing to the society.

Step 4. Determine the criteria or standards that will be used to assess the value, correctness, or desirability of the behaviour elicited. In simple paper-and-pencil tests, this step consists of determining the 'scoring key', i.e., the list of correct answers. In more complex or non-symbolic situations, such as those in which a real-life performance is to be evaluated - e.g., a dance, a lathe operation, a meal preparation, or a discussion-group participation - this step consists of (a) a list of important dimensions of the performance and (b) an accompanying set of scales for use by observers, measurers, or judges of the performance.

Some samples of items from tests involving non-symbolic situations are shown in Figures 1 and 2.

The student is shown a setup consisting of a hack saw, a vice holding a piece of work, and a saw cut started. He is asked, "Which one of the following is a correct statement?"

(a) The hack saw blade is in the frame backwards.

(b) A finer toothed blade should be used on this job.

(c) The piece protrudes too far from the vice.

(d) One should saw on the other side of the vice.

(Note that any of the four alternative responses could be made the desired response.)

The student is shown a lathe setup with a metal cylinder in place. The cylinder has been roughly cut. He is asked, "Which one of the following statements explains why this lathe tool bit cut rough?"

(a) The feed is too high.

(b) The speed is too fast.

(c) The work is too loose between centres.

(d) The cutting edge is too high.

(Again, any of the four alternative responses could be made the correct one.)

Figure 1. Test situations for a shop course

Needless to say, the criteria to be used should be determined by the values, customs, and standards prevalent in the culture of the society in whose education system the evaluation is being performed. Tests, criteria, and standards appropriate in one culture will in many instances be inappropriate in another. Only curriculum experts in any particular country can judge and ensure this appropriateness of criteria and standards.

Food Score Card

Plantain

Appearance			Score
1	2	3	2
Shrivelled		Slightly Moist	

Colour			Score
1	2	3	3
Pale or Burned		Well-Browned	

Moisture Content			Score
1	2	3	1
Dry		Juicy	

Taste			Score
1	2	3	1
Raw, Tasteless or Burned		Flavour Developed	

Figure 2. A rating scale for evaluating a food product in a home economics course. The dimensions of cooked plantain to be evaluated are its appearance, colour, moisture content and taste. For each of these a three-point rating scale is provided.

Step 5. Apply the measuring instruments - the tests, rating scales, situational performance tests, etc. - to the students whose achievement of the objectives is being determined. This step refers to the administration of the test. It can take the form of individual or group testing under highly or loosely standardized conditions, depending on what is necessary and feasible for the purposes of the evaluation.

Individual testing is, of course, much more expensive than group testing. It requires much more time on the part of the examiners, and frequently it is more important in individual testing that the examiner be highly trained. Group tests, by definition, can be given to groups ranging in size from 2 to thousands of students at a time, depending on the size of the auditorium and the number of proctors available. The cost of testing each student is thereby reduced materially.

Tests that require judgmental scoring by experts are expensive. Tests that can be scored objectively by clerks or by machines are much less expensive.

Thus, in general, group and objectively scorable testing is to be preferred except under special conditions. Individual tests can take the form of oral examinations and interviews, and their high expense is justifiable when the nature of the achievement to be evaluated permits no valid alternative. Examples would be tests of pronunciation or certain problem-solving processes. These conditions also include those in which the testing must be carefully adjusted to the child or adult being tested. If highly idiosyncratic behaviours are being looked for -- behaviours of the kind that might be exhibited by exceptional children or adults (either retarded or highly gifted), or are to be examinea for signs of creativity and originality -- then individual testing may be necessary. But, for the vast majority of educational objectives for the vast majority of students, group tests will serve. With sufficient technical skill, such tests, even when concerned with complex cognitive processes, can also be made objectively scorable.

Step 6. Score the behaviour - the responses or performances - elicited by the measuring instruments. The score will take such forms as 'number of correct responses' or 'total rating on the various dimensions' or 'accuracy of the product prepared in relation to specified attributes'. The score should tell the degree to which the objectives embodied in the questions or items of the measuring instruments have been achieved by the student.

Step 7. Judge the degree to which the score obtained by the students reflects achievement of the objectives of the educational activity or enterprise. Two major approaches to making such judgements are termed the norm-referenced and the criterion-referenced approaches. In the former, each student's score is compared with that of other students constituting a norm group. Such a norm group might be 'other students who have just taken the same course', or it might be 'a representative sample of students of the same grade-level throughout the nation'. The given student is then found to have a certain percentile rank in relation to the norm group, i.e., to equal or exceed a certain percentage of the students in the norm group.

The success of a norm-referenced test is judged in part by the degree to which it discriminates among students. The purpose of the test is to spread students out - to put them in a rank order ranging from the highest achieving to the lowest. A test that does not discriminate in this way is judged to be unsuccessful in reflecting variance, and hence it cannot be an effective test, if norm-referenced approaches are being used. Norm-referenced measurement, when used to assess the achievement of individual students, imposes a kind of competition among students. Each student's achievement is evaluated primarily by being compared with that of the other students in his class, school, or community. By this approach, some students are forced to be inferior, because some students must by definition be below average; indeed, half of the students must always fall below the median, of course, and suffer the corresponding implications of inferiority.

For many years, it was considered necessary to use the norm-referenced approach because it was assumed that no other way of judging educational achievement - other than by comparing that of one student with that of other students - was possible. More recently, it has been realized that achievement measurement can be referred to objectives rather than to other students. If it is possible to set up educational or instructional objectives, and to define observable behaviours that indicate achievement of those objectives, then it is possible to evaluate achievement by eliciting and evaluating those behaviours directly, without reference to the performance of other students.

In criterion-referenced interpretation of test scores, the judgement is made directly with reference to the pre-specified types of behavioural or performance objectives of the educational activity. If the objectives referred to 'ability to read a typical newspaper article', then the interpretation of the student's score refers directly to whether such an ability has been demonstrated. If the objective referred to 'ability to produce a steel cylinder', with dimensions of a certain accuracy, on a lathe, then the criterion-referenced interpretation refers directly to whether such an ability has been demonstrated.

Other examples could be found in the objectives of programmes for training vehicle drivers, or subsistence farmers, etc. So-called 'minimal learning packages' intended to develop basic skills can be evaluated by criterion-referenced approaches.

Norm-referenced evaluation has been used frequently in the past. It is closely linked to 'grading on the curve', which requires that only certain percentages of students be regarded as excellent, good, fair, etc. Criterion-referenced evaluation is coming to be used more and more as newer approaches to education are adopted - approaches such as programmed instruction, individualized instruction, self-paced instruction, and mastery-learning approaches.

In criterion-referenced measurement, the shape of the ideal or desired distribution of test scores is not the traditional 'bell-shaped' or any other distribution that shows a great amount of variance among students. Rather, the ideal distribution is one that shows the maximum possible number of students achieving the maximum or almost-maximum number of instructional objectives. That is, in criterion-referenced measurement one hopes that scores will pile up at the high end of the distribution. To the degree that such a goal is not achieved, one looks toward improving methods of instruction rather than resigning oneself to accepting inevitable individual (and perhaps hereditary) differences in aptitude or intelligence among students.

The basic distinction between norm-referenced and criterion-referenced measurement has implications for various other aspects of achievement evaluation. As already indicated, in criterion-referenced testing the variability yielded by the test is to be minimized, while in norm-referenced testing it is to be maximized. Similarly, in criterion-referenced testing the main criterion is relevance to, or significance for, educational objectives, or the criterion of successful educational achievement, regardless of either the 'difficulty' or the degree of 'discrimination' achieved by the test question or item. Similarly, the reliability of a test (i.e. consistency of measures obtained with it), insofar as it depends on the test's producing variability between students, may become irrelevant. If all of the students get perfect or nearly perfect scores, the test will have low reliability in the usual senses that depend on the variance among test scores, and yet be highly reliable and valid in the sense that the student's achievement consistently reflects his achievement of the objectives.

THE FORMATIVE AND SUMMATIVE FUNCTIONS OF EVALUATION

It is regarded as axiomatic that one of the major purposes of any evaluation project is to improve the quality of education in the country in which the project is operating. Contrary to what is sometimes thought, evaluation is not an end in itself ; its main purpose is not merely to provide information on whether or not a particular project has achieved the kind of success intended. Rather, in the long run, evaluation should help those responsible for the development and execution of the project to ensure that success is actually achieved. There are, however, numerous occasions when circumstances will permit only the assessment of final success, and such evaluation also can often produce information of value.

Misunderstandings about the function of evaluation arise from the failure to distinguish between summative and formative evaluation. The former comes at the end of an educational activity or project, while the latter occurs continually during the activity or project, with the aim of producing information which can be fed back to ensure that the appropriate aims and objectives are being attained. The information fed back from formative evaluation may necessitate changes in any of

the first four steps listed earlier. They may call for changes in the situation or procedures developed in order that specific objectives will be achieved, or changes in the criteria or standards used to assess the value of the required behaviour. There are occasions, however, when modifications may be required in the objectives themselves.

In theory, whether an evaluation is to be summative or formative, the objectives to be achieved should be clearly stated at the outset of a new project. In practice, it is often the case that some summative evaluations are added on to a project almost as an afterthought. In such instances, formulations of objectives may not be produced until the project is almost completed. The danger here is that the evaluation may then be concerned with objectives which have in fact not actually been operative, rather than with those which the project was actually intended to achieve.

The uses of summative evaluation

Two kinds of circumstances can be identified in which summative evaluation may be desirable. The first arises in a project, usually a short term one, in which the circumstances will hardly permit of any changes being introduced during the course of the project. An example would be a project which required building a number of additional secondary schools, or further teacher-training colleges, without any initial specifications that the curricula in the new establishments should be changed to meet new requirements. The summative evaluation would then be largely concerned with quantitative aspects, or perhaps with determining whether there were differences between the average levels of achievement of students in the new institutions and those of students in already existing institutions.

In making a summative evaluation, it is often desirable to have a baseline against which the results of a given educational activity can be measured. Such a baseline can be obtained from the performance of a 'control group' - a group of students who have not received the kind of educational experience or training being evaluated. Sometimes such a control group received no training at all. In other instances, the control group may receive the traditional or regular kind of training that one is trying to improve upon. Ideally, the experimental group - the one that receives the new kind of training - and the control group - the one that receives the old kind of training or none at all - are 'randomly equivalent'. That is, students in these two groups are assigned at random from the total group, by some mechanism such as a table of random numbers or the tossing of a coin. Such randomization, if it is feasible, and if the control and experimental groups are large enough, is sufficient to ensure that all other possible explanations of any significant post-training differences between the two groups cannot be attributed to extraneous factors, such as differences in aptitude, age, social class, home background, or whatever. Randomization makes the two groups non-significantly different in all conceivable factors other than the experimental variable - the difference between the new and old kinds of education, curriculum, or instruction.

Sometimes it is possible to introduce additional refinements by using statistical methods to adjust for whatever differences between the experimental and control groups may remain even after random assignment. Thus, if the two groups are found to differ somewhat (even if only to a chance degree) in scholastic aptitude, even after randomization, the effect of this difference on the post-instructional achievement test scores can be controlled statistically. If the adjusted achievement scores still differ significantly, i.e., to a degree greater than can be accounted for by chance fluctuations in random sampling, then it must be inferred that the instructional differences made a genuine difference in achievement that cannot be attributed to differences in aptitude.

The second circumstance in which summative evaluation may be desirable occurs at the end of a project during which formative evaluation had been carried out. The primary aim of formative evaluation is to ensure, through changes brought about by the feedback process, that the originally stated objectives of a project are being achieved. In many instances, however, the feedback process causes changes to be made in the initially specified objectives, because some may prove to be unattainable in their original form. Where this happens, a summative evaluation should report on the extent to which both original and modified objectives have been achieved. An instance of this second type of summative evaluation might occur in a project which required, say, "revisions of the secondary school structure and curriculum in order to improve the standard of secondary school education".

It has already been emphasized that the translation of general goals into appropriate sets of behavioural objectives forms the first important step in the evaluation process. Detailed behavioural objectives take time to be developed, however, even by an expert. It is not reasonable to expect that in the time available they will readily flow from the pen of a member of a Project Identification Mission, for example, whatever expertise in this area he may possess. It is suggested that at this stage and in the appraisal reports much more detail about intended qualitative objectives be included. Then a set of detailed behavioural objectives can be produced by the time a project is ready to get under way.

The uses of formative evaluation

Since the major aim of formative evaluation is to improve education, the collection of data on the achievement of any objective should be undertaken as soon as appropriate after a project begins. The results of this assessment should be fed back immediately to those concerned with the project's development. If these results show that the stated objectives are indeed being achieved, the project can continue along the lines already adopted. If results reveal, however, that the improvement sought, or the achievement of the standards desired, is falling below expectation, then steps should be taken to change the project in a direction that subsequent evaluation will reveal to be more appropriate for the achievement of the stated objectives. Only continual evaluation of this kind can ensure that the objectives will be achieved. Such formative evaluation does, in fact, occur in the construction of hardware (e.g., school buildings) when architects inspect and modify construction projects in process. It is here being recommended that it also occur as a matter of routine in the conduct of educational programmes and courses.

It will be seen from the above that the usefulness of summative evaluation is somewhat limited. It is rarely possible, with such evaluation, to make judgements or conclusions about any improvement that may have taken place since the inception of a project. Further, there is no opportunity to effect a change if the summative evaluation demonstrates that particular objectives have not been achieved.

This brief introduction to the process of evaluation implies a number of important questions about the evaluation of projects already under way, but not yet completed. It also implies the need to consider the whole question of evaluation at the very outset of new projects.