evaluation planning 2

evaluation planning

This second chapter presents basic concepts in the field of educational *evaluation*. It stresses the very close relationship between evaluation and definition of educational objectives; and the primary role of any evaluation, which is to facilitate decision-making by those responsible for an educational system. It defines the subject, the purpose, the goals and the stages of evaluation and highlights the concepts of validity and relevance

Those who would like to learn more about these problems should consult the following publications

- Public Health Papers No. 52, WHO, "Development of educational programmes for the health professions", 1974
- Evaluation of school performance, Educational documentation and information. Bulletin of the International Bureau of Education No. 184, third quarter 1972, 84 pages

After having studied this chapter and the reference documents mentioned you should be able to:

1. Draw a diagram showing the relationship between evaluation and the other parts of the educational process.

2. Define the principal role of evaluation, its purpose and its aims.

3. Describe the difference between formative and certifying evaluation.

4. List the good and bad features of a test.

5. Compare the advantages and disadvantages of tests in current use

6. Define the following terms validity, reliability, objectivity, and describe the relationship that exists between them.

7 Choose an appropriate evaluation method (questionnaire, written examination, "objective" test (MCQ or short answer questions) or essay question, oral examination, direct observation, etc.) for measuring the students' attainment of a specific educational objective. Compare the alternatives in a specification table.

8. Define (in the form of an organizational diagram) the organization of an evaluation system suitable for your establishment, and list the stages involved

Indicate

(a) the most important educational decisions you have to take;

(b) the data to be collected to provide a basis for those decisions;

(c) the aims of the system and sub-systems in terms of decisions to be taken and the object of each decision (teachers, students, programmes)

9 Identify obstacles to and strategies for improvement of a system of evaluating students, teachers and programmes.

□ □

To change curricula or instructional methods without changing examinations would achieve nothing!

Changing the examination system without changing the curriculum had a much more profound impact upon the nature of learning than changing the curriculum without altering the examination system

G.E. Miller*

□ □

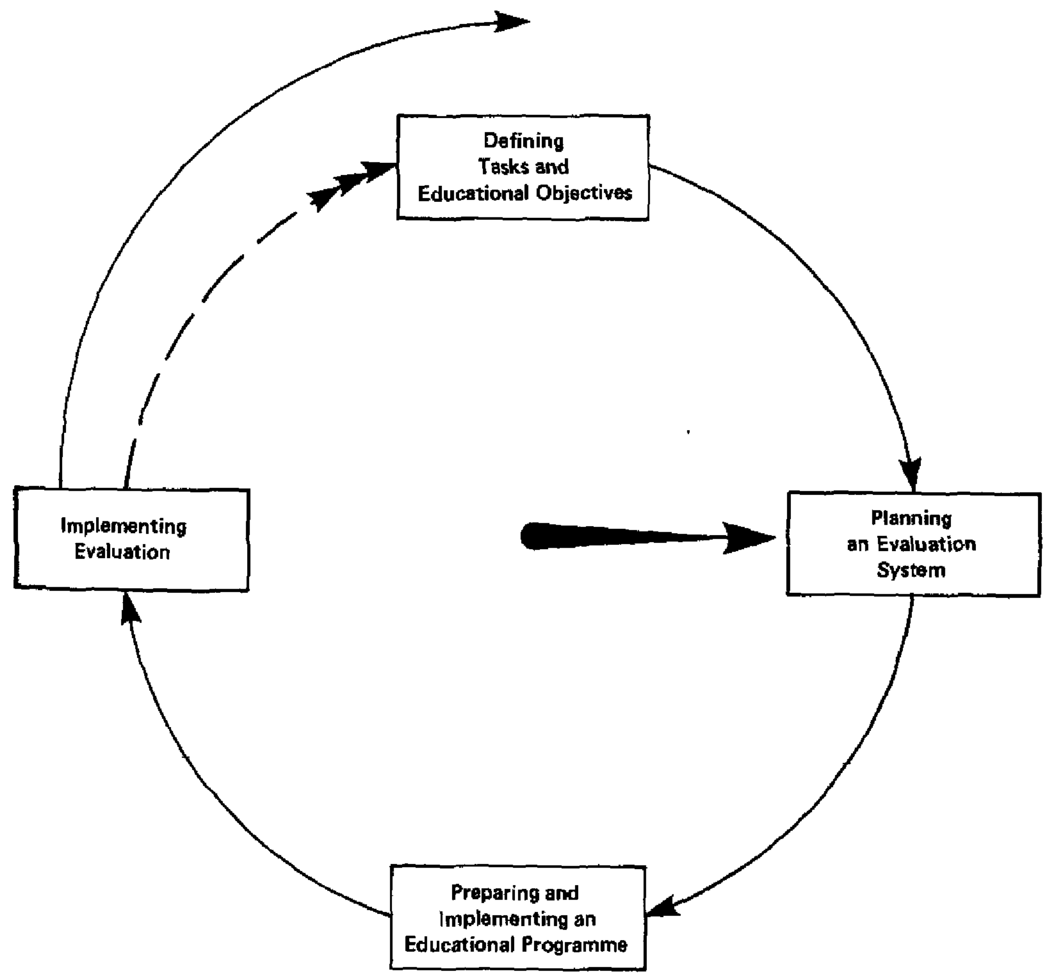
* International Medical Symposium No. 2. Rome. 23–26 March 1977

An analysis of educational innovations all over the world confirms G. Miller's opinion In this second chapter, therefore, you are invited to *plan* a system of evaluation that can be used as a basis for preparation and implementation of a programme. The process is already under way, for the formulation of specific educational objectives requires definition of criteria indicating the minimum level of performance expected from the student. Educational decisions have to be made frequently during preparation and implementation of a programme; and the main purpose of evaluation is in fact to provide a basis for value judgments that permit better educational decision-making. First of all you must decide what you want to evaluate: students, teachers and/or programmes. In each case you must determine what important educational decisions you will be expected to make in your capacity as teacher or administrator, for the instruments and mechanisms of evalution providing data for value judgments will be developed and used according to the type of decision required. A general methodology of evaluation and corresponding techniques do exist. Some are simple; others very complex and costly in time and money. Here again you will make your choice according to criteria that will ensure an adequate level of security As in every educational process, you will have to shape all the consequences of your decisions into a coherent and logical whole You are therefore invited to read the next pages before doing the exercise on p 2 09.

The person who sets the examination controls the programme.

Education by objectives is not possible unless examinations are constructed to measure attainment of those objectives.

the educational spiral



The evaluation process provides a basis for value judgments that permit better educational decision-making

Notice to all teachers

You are reminded that evaluation of education must begin with a clear and meaningful definition of its objectives

Evaluation . . . . .of whom?

. . . . . .of what?

- **Students**
- **Teachers**
- **Programmes and courses**

. . . . in relation to what?

- **In relation to educational objectives.**
  (They are the common denominator.)

EXERCISE

Answer question 2 on p. 2.45.
Check your answer on p. 2.48.

EXERCISE

Before starting to define the organization, stages or methods of an *evaluation* system suitable for the establishment in which you are teaching, it would be useful to state:

What important educational *decisions** you think you and your colleagues will be taking over the next three years.

* Examples of educational decisions.

- to decide which students will be allowed to move up from the first to the second year
- or to decide to purchase an overhead projector rather than a blackboard
- or to decide to appoint Mr. X full professor.

or to decide. . . . .

You and your colleagues will have to make value judgments as a basis for each decision. It will therefore be useful to plan the construction and use of "instruments of evaluation" that will enable you to collect the data needed for making those value judgments

Personal Notes

evaluation
— a few assumptions*

- Education is a process, the chief goal of which is to bring about changes in human behaviour
- The sorts of behavioural changes that the school attempts to bring about constitute its objectives
- **Evaluation consists of finding out the extent to which each and every one of these objectives has been attained, and the quality of teaching techniques and teachers.**

* adapted from Downie

assumptions underlying basic educational measurement and evaluation*

Human behaviour is so complex that it cannot be described or summarized in a single score.

The manner in which an individual organizes his behaviour patterns is an important aspect to be appraised. Information gathered as a result of measurement or evaluation activities must be interpreted as a part of the whole. Interpretation of small bits of behaviour as they stand alone is of little real meaning.

The techniques of measurement and evaluation are not limited to the usual paper-and-pencil tests Any bit of valid evidence that helps a professor or counsellor in better understanding a student and that leads to helping the student to understand himself better is to be considered worthwhile. Attempts should be made to obtain all such evidence by any means that seem to work

The nature of the measurement and appraisal techniques used influences the type of learning that goes on in a classroom If students are constantly evaluated on knowledge of subject-matter content, they will tend to study this alone Professors will also concentrate their teaching efforts upon this. A wide range of evaluation activities covering various objectives of a course will lead to varied learning and teaching experiences within a course

The development of any evaluation programme is the responsibility of the professors, the school administrators, and the students. Maximum value can be derived from the participation of all concerned.

the philosophy of evaluation*

1. Each individual should receive that education that most fully allows him to develop his potential.

2. Each individual should be so placed that he contributes to society and receives personal satisfaction in so doing.

3. Fullest development of the individual requires recognition of his essential individuality along with some rational appraisal by himself and others

4. The judgments required in assessing an individual's potential are complex in their composition, difficult to make, and filled with error.

5. Such error can be reduced but never eliminated. Hence any evaluation can never be considered final.

6. Composite assessment by a group of individuals is much less likely to be in error than assessment made by a single person.

7. ~~The efforts of a conscientious~~ group of individuals to develop more reliable and valid appraisal methods lead to the clarification of the criteria for judgment and reduce the error and resulting wrongs.

8. Every form of appraisal will have critics, which is a spur to change and improvement

* Downie, N M *Fundamentals of Measurement Techniques and Practices* New York Oxford University Press, 1967

evaluation is

a continuous process

based upon criteria

cooperatively developed

concerned with measurement of the performance of learners, the effectiveness of teachers and the quality of the programme*

* This chapter is mainly concerned with the evaluation of students. Evaluation of programmes and teachers is dealt with in chapter 4.

the psychology of evaluation*

1. For evaluation activities to be most effective, they should consist of the best possible techniques, used in accordance with what we know to be the best and most effective psychological principles.

2. For many years readiness has been recognized as a very important prerequisite for learning. A student is ready when he understands and accepts the values and objectives involved.

3. It has long been known that people tend to carry on those activities which have success associated with their results. This has been known as Thorndike's ***Law of Effect.*** Students in any classroom soon come to realize that certain types of behaviour are associated with success – in this case, high marks on a test or grades in a course. Thus, if a certain teacher uses tests that demand rote memory, the students will become memorizers. If a test, on the other hand, requires students to apply principles, interpret data, or solve problems, the students will study with the idea of becoming best fitted to do well on these types of test items. ***In the long run, the type of evaluation device used determines, to a great extent, the type of learning activity in which students will engage in the classroom***

4. Early experiments in human learning showed that individuals learn better when they are "constantly" appraised in a meaningful manner as to how well they are doing.

5. The ***motivation*** of students is one of the most important – and sometimes the most difficult to handle – of all problems related to evaluation. It is redundant for us to say that a person's performance on a test is directly related to his motivation. Research has shown that when a student is really motivated, performance is much closer to his top performance than when motivation is lacking.

6. ***Learning is most efficient when there is activity on the part of the learner***

EXERCISE

Try to answer question 3 on p. 2.45

Check your answer on p. 2.48.

*Downie N M, *Fundamentals of Measurement Techniques and Practices* New York Oxford University Press, 1967

continuous evaluation
formative and certifying evaluation

Evaluation of education must begin with a clear and meaningful definition of its objectives We cannot measure something unless we have first defined what it is we wish to measure.

When this phase of evaluation (the definition of objectives) has been properly completed, the choice or development of suitable examination procedures is that much easier Schematically represented, the education cycle (p. 2.05) comprises the determination of objectives, the planning of an evaluation system, the development of teaching activities and the implementation of evaluation procedures with possible revision of objectives

The role of evaluation should not be limited to one of penalization It should not be just a series of only too frequent obstacles which the students are supposed to get over and which become their sole subject of concern, the actual instruction becoming quite secondary. Under these circumstances the student's only interest is how to obtain his diploma with least effort It is the teacher's responsibility to convince the student that his education is directed towards wider aims than merely gaining a diploma and that helping him to do so is not the sole purpose of evaluation (see p. 2.18)

Evaluation should also be *formative*, providing the student with information on his progress It must therefore be continuous. This concept has often been misinterpreted, resulting in constant harassment of the student. There is a fundamental difference between *formative* and *certifying* evaluation

You will find the following equivalents in the literature for these two expressions

Formative evaluation	Certifying evaluation
or	or
diagnostic evaluation	summative evaluation

Continuous evaluation must pit the student against himself and his own lack of knowledge and not against other students.

formative evaluation*

- is designed to inform the student about the amount he still has to learn before achieving his educational objectives;
- measures the *progress* or *gains* made by the student from the moment he begins a programme until the time he completes it,
- enables learning activities to be adjusted in accordance with progress made or lack of it,
- *should in no way be used by the teacher to make a certifying judgment,* the anonymity of the student should be safeguarded by use of a code of his choice. The coding system makes it possible to follow the progress of individuals and groups while preserving anonymity
- is controlled in its use by the student (results should not appear in any official record);
- is very useful in guiding the student and prompting him to ask for help;
- is carried out frequently – as often as the student feels necessary;
- provides the teacher with qualitative and quantitative data for modification of his teaching (particularly contributory educational objectives) or otherwise

certifying evaluation

- is designed to protect society by preventing incompetent personnel from practising.
- is traditionally used for placing students in order of merit and justifying decisions as to whether they should move up to the next class or be awarded a diploma
- is carried out less frequently than formative evaluation. usually at the end of a unit or period of instruction.

EXERCISE

Try to answer questions 4 – 8 on p. 2.45.

Check your answers p. 2.48.

We don't care how hard the student tried, we don't care how close he got . . until he can perform he must not be certified as being able to perform.

R.F Mager

* Read the article by C. McGuire – Diagnostic examinations in medical education. In: Public Health Papers No. 52, WHO, Geneva, 1973.

Personal Notes

Strict Rule

Formative evaluation should in no way be used by the teacher against the student.

student evaluation : what for

9 Incentive to Learn (motivation)

10 Feedback to student

11 Modification of Learning Activities

12 Selection of students

13 Success or failure

14 Feedback to teacher

15 School public relations

16 Protection of society (certification of competence)

$\Bigg\}$ appropriate measuring techniques

The numbers on the left refer to the exercise on the next page and the questions on pp. 2.45 – 2.47

aims of student evaluation*

1 To determine success or failure on the part of the student. This is the conventional role of examinations (certifying evaluation).

2. To provide "feedback" for the student: to keep him constantly informed about the instruction he is receiving, to tell him what level he has reached, and to make him aware through the questions of what parts of the course he has not understood (formative and certifying evaluation).

3. To provide "feedback" for the teacher to inform him whether a group of students has not understood what he has been trying to explain. This enables him to modify his teaching where necessary to ensure that what he wishes to communicate to the students is correctly understood (formative and certifying evaluation)

4 The "reputation of the school" is something of which the importance is not always evident, at least in European schools, whose reputation is often based not on an examination system but on long-standing traditions. North American schools, on the other hand, customarily publish the percentage of students who have passed, for example, national examinations (formative and certifying evaluation).

EXERCISE

Now try it . . indicate for each of the aims of evaluation (numbered 9–16 on the previous page) whether the measurement technique will be of the Certifying Evaluation type (C) or both Certifying *and* Formative Evaluation (CF). Check your answers on p. 2.48

* Adapted from Downie, N M *Fundamentals of Measurement Tecnniques and Practices* New York Oxford University Press 1967

four steps in student evaluation

taking as one basis

The **criteria** (acceptable level of performance) **of the educational objectives**

Development and use of measuring **instruments**

Interpretation of measurement data

Formulation of **judgments** and taking of appropriate **action**

common methodology for student evaluation

- Evaluation of practical skills
- Evaluation of communication skills
- Evaluation of knowledge and intellectual skills

1. Make a list of observable types of behaviour showing that the objective pursued *has been reached*

2 Make a list of observable types of behaviour showing that the objective pursued has *not* been reached.

3 Determine the *essential* features of behaviour in both lists.

4 Assign a positive or negative *weight* to the items on both lists.

5 Decide on the *acceptable performance* score.

* For the last three stages obtain the agreement of several experts.

Example. *Objective*. Reassure the mother of a child admitted to hospital

Attitude	-2	-1	0	+1	+2
Explain clearly what has been done to the child	often uses medical terms and never explains what they mean	often uses medical terms and rarely explains what they mean	rarely uses medical terms and sometimes explains what they mean	rarely uses medical terms and always explains what they mean	uses only terms suited to the mother's vocabulary

etc. See the complete table on p 4 24.

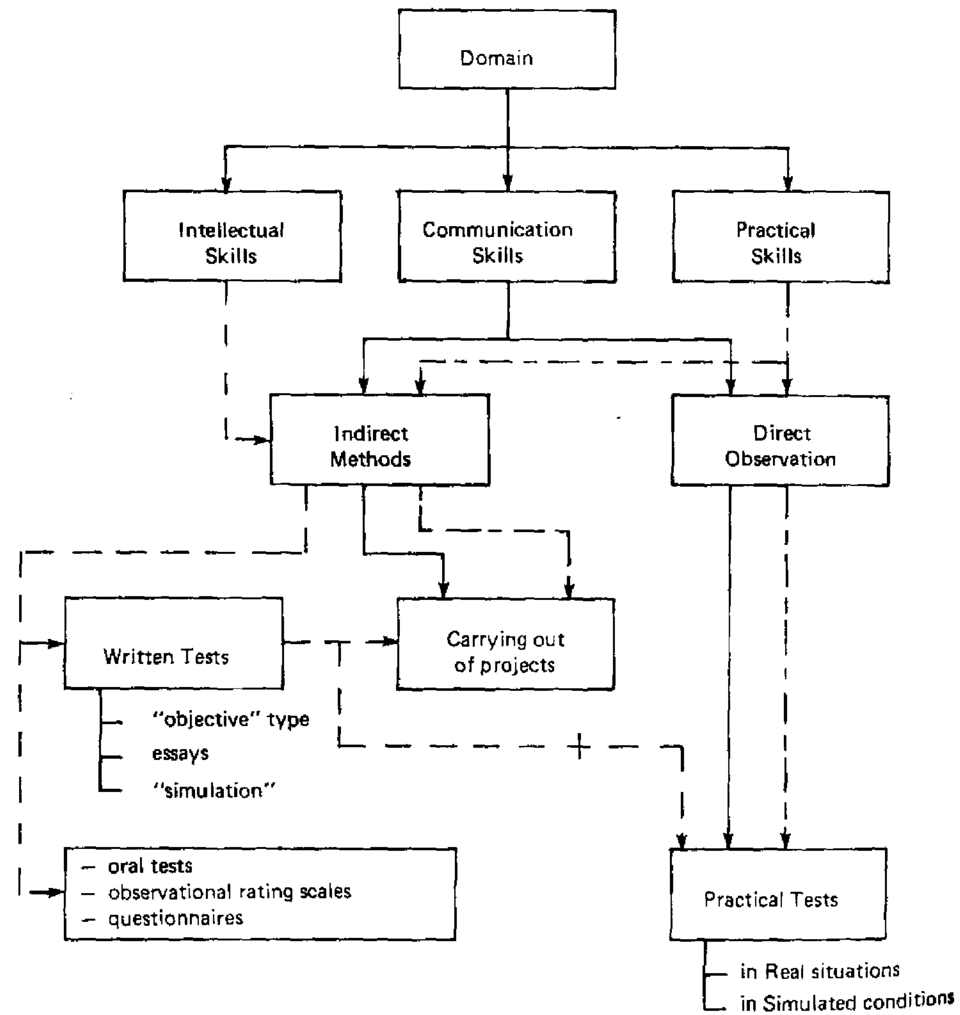
Minimum Performance Score The student should score n marks out of 10 on the rating scale

EXERCISE

Try to answer questions 17 – 20 on p. 2 46 and check your answers on p. 2.48.

See also Rezler, Public Health Papers No 52, pp 70 – 83

evaluation methodology according to domains to be evaluated



EXERCISE

For each of the educational objectives you have already defined (pp. 1.54, 1.55), choose from among the methods of evaluation set out on p.2.22 the one you think most suitable for informing you and the student* on the extent to which the objective has been achieved.

	Objective	Method of evaluation	Instrument of evaluation
EXAMPLES	1 page 1.50	Indirect method	Short, open answer question based on observation of a patient
	2 page 1.50	Indirect method	Questionnaire
	3 page 1.50	Direct observation	Practical examination
1			
2			
3			
4			
5			

* For the purposes of this exercise the total number of students to be considered should be fixed: e.g., 100.

Personal Notes

general remarks concerning examinations

Analysis of the most commonly used tests shows that sometimes, often even, the questions set are ambiguous, unclear, disputable, esoteric or trivial It is essential for anyone constructing an examination, whether of the traditional written type, an objective test or a practical test, ***to submit it to his colleagues for criticism*** to make sure that its content is relevant (related to an educational objective) and of general interest, and does not exclusively concern a special interest or taste of the author, that the subject is interesting and real for the general practitioner or the physicians with a specialty different from that of the author, and that the questions (and the answers in the case of multiple-choice questions) are so formulated that experts can agree on the correct response. It is clear that a critical analysis along these lines would avoid the over-simplification of many tests which only too often justifies the conclusion· "the more you know about a question the lower will be your score"

The author of a test ***is not*** the best judge of its clarity, precision, relevance and interest. ***Critical review of the test by colleagues is consequently essential for its sound construction***

Moreover, an examination must take the factor of ***practicability*** into account This will be governed by the ***time necessary*** for its construction and administration, scoring and interpretation of the results, and by its general ease of use.

If the examination methods employed become a burden on the teacher because of their impractical nature he will tend not to assign to the measuring instrument the importance it deserves.

□ □

A discussion is not always pertinent to the problem at hand, but one learns to allow for some rambling It seems to help people realize that they normally use quite a few fuzzies during what they consider "technical discussions", it helps them realize that they don't really know what they are talking about . . a little rambling helps clear the air. Asking someone to define his goal in terms of performance is a little like asking someone to take his clothes off in public – if he hasn't done it before, he may need time to get used to the idea.

R.F. Mager

□ □

qualities of a test

- **Directly related to educational objectives**
- **Realistic & practical**
- **Concerned with important & useful matters**
- **Comprehensive but brief**
- **Precise & clear**

considerations of the type of competence a test purports to measure

No test format (objective, essay or oral) has a monopoly on the measurement of the highest and more complex intellectual processes Studies of various types of tests support the view that the essay and the oral examination, as commonly employed, test predominantly simple recall and, like the objective tests in current use, rarely require the student to engage in reasoning and problem-solving In short, the *form of a question does not determine the nature of the intellectual process required to answer it.*

Second, there is often a tendency to confuse the difficulty of a question with the complexity of the intellectual process measured by it However, it should be noted that a question requiring simple recall may be very "difficult" because of the esoteric nature of the information demanded; alternatively, a question requiring interpretation of data or application of principles could be quite "easy" because the principles of interpretation are so familiar and the data to be analysed so simple. In short, *question difficulty and complexity of instructions are not necessarily related to the nature of the intellectual process being tested.*

Third, there is often a strong inclination to assume that any question which includes data about a specific case necessarily involves problem-solving, whereas, in fact, "data" are often merely "window dressing" when the question is really addressed to a general condition and can be answered equally well without reference to the data Or, the data furnished about a "specific case" may constitute a "cut-and-dried", classical textbook picture that, for example, simply requires the student to recall symptoms associated with a specific diagnosis. It is interesting to note that questions of this type can readily be converted into problems that do require interpretation of data and evaluation, simply by making the case material conform more closely to the kind of *reality* that an *actual* case, rather than a textbook, presents.

In short, just as each patient in the ward or out-patient department represents a unique configuration of findings that must be analysed, *a test which purports to measure the student's clinical judgment and his ability to solve clinical problems must simulate reality* as closely as possible by presenting him with specific constellations of data that are in some respects unique and, in that sense, are new to him. Do not try to use a MCQ or a SOAQ to find out whether the student is able to communicate orally with a patient!

However reliable or objective a test may be, it is of no value if it does not measure ability to perform the tasks expected of a health worker in his/her professional capacity.

common defects of examinations (domain of intellectual skills)

A review of examinations currently in use strongly suggests that the most common defects of testing are:

triviality the triviality of the questions asked, which is all the more serious in that examination questions can only represent a small sample of all those that could be asked. Consequently it is essential for each question to be important and useful;

outright error outright error in phrasing the question (or, in the case of multiple-choice questions, in phrasing the distractors and the correct response),

ambiguity ambiguity in the use of language which may lead the student to spend more time in trying to understand the question than in answering it; in addition to the risk of his giving an irrelevant answer,

conservatism forcing the student to answer in terms of the bias or even the outmoded ideas of the examiner, a bias which is well known and often aggravated by the teaching methods themselves (particularly traditional lectures),

complexity complexity or ambiguity of the subject matter taught, so that the search for the correct answer is more difficult than was anticipated;

unintended cues unintended cues in the formulation of the questions that make the correct answer obvious; this fault, which is often found in multiple-choice questions, is just as frequent in oral examinations.

outside factors to be avoided

In constructing an examination, outside factors must *not* be allowed to interfere with the factor to be measured.

complicated instructions (ability to understand instructions) In some tests, the instructions for students on how to solve the problems are so complicated that what is really evaluated is the students' aptitude to understand the question rather than their actual knowledge and ability to use it. This criticism is often made of multiple-choice examinations in which the instructions appear too complicated. The complexity is often more apparent than real and disturbs the teacher rather than the student.

over-elaborate style (ability to use words) The student may disguise his lack of knowledge in such elegant prose that be succeeds in influencing the corrector, who judges the words and style rather than the student's knowledge.

trap questions (ability to avoid traps) This type of interference does not depend on a measuring instrument, but on possible sadistic tendencies on the part of the examiner who, during an examination, may allow himself to be influenced by the candidate's appearance, sex, etc. Some candidates are more or less skilled at playing on these tendencies.

"test-wise" This is a criticism that is generally made of multiple-choice examinations, it may in fact be applied to other forms of evaluation. In oral and written examinations, students develop a sixth sense, often based on statistical analysis of past questions, which enables them somehow to predict the questions that will be set.

comparison of advantages and disadvantages of different types of test

Oral examinations

Advantages

1. Provide direct personal contact with candidates.
2. Provide opportunity to take mitigating circumstances into account.
3. Provide flexibility in moving from candidate's strong points to weak areas.
4. Require the candidate to formulate his own replies without cues.
5. Provide opportunity to question the candidate about how he arrived at an answer.
6. Provide opportunity for simultaneous assessment by two examiners.

Unfortunately all these advantages are rarely used in practice.

Disadvantages

1. Lack standardization.
2. Lack objectivity and reproducibility of results.
3. Permit favouritism and possible abuse of the personal contact.
4. Suffer from undue influence of irrelevant factors.
5. Suffer from shortage of trained examiners to administer the examination.
6. Are excessively costly in terms of professional time in relation to the limited value of the information it yields.

Practical examinations

Advantages

1. Provide opportunity to test in a realistic setting skills involving all the senses while the examiner observes and checks performance.
2. Provide opportunity to confront the candidate with problems he has not met before both in the laboratory and at the bedside, to test his investigative ability as opposed to his ability to apply ready-made "recipes".
3. Provide opportunity to observe and test attitudes and responsiveness to a complex situation (videotape recording).
4. Provide opportunity to test the ability to communicate under pressure, to discriminate between important and trivial issues, to arrange the data in a final form.

Disadvantages

1. Lack standardized conditions in laboratory experiments using animals, in surveys in the community or in bedside examinations with patients of varying degrees of cooperativeness*
2. Lack objectivity and suffer from intrusion or irrelevant factors.
3. Are of limited feasibility for large groups.
4. Entail difficulties in arranging for examiners to observe candidates demonstrating the skills to be tested.

*Standardized practical tests can be constructed; see "Simulation and evaluation in medicine" in Public Health Papers No. 61, pp.18–34.

Essay examinations

Advantages

1. Provide candidate with opportunity to demonstrate his knowledge and his ability to organize ideas and express them effectively.

Disadvantages

1. Limit severely the area of the student's total work that can be sampled.
2. Lack objectivity.
3. Provide little useful feedback.
4. Take a long time to score.

Multiple-choice questions

Advantages

1. Ensure objectivity, reliability and validity; preparation of questions with colleagues provides constructive criticism.
2. Increase significantly the range and variety of facts that can be sampled in a given time.
3. Provide precise and unambiguous measurement of the higher intellectual processes.
4. Provide detailed feedback for both student and teachers.
5. Are easy and rapid to score.

Disadvantages

1. Take a long time to construct in order to avoid arbitrary and ambiguous questions.
2. Also require careful preparation to avoid preponderance of questions testing only recall.
3. Provide cues that do not exist in practice.
4. Are "costly" where number of students is small.

It is a highly questionable practice to label someone as having achieved a goal when you don't even know what you would take as evidence of achievement.

R.F. Mager

Personal Notes

evaluation in education
qualities of a measuring instrument

1 *Some definitions*

1 1 It should be recalled once more that ***education*** is defined as a process developed for bringing about changes in the student's behaviour. At the end of a given learning period there should be a greater probability that types of behaviour regarded as ***desirable*** will appear; other types of behaviour regarded as undesirable should disappear

1.2 The ***educational objectives*** define the desired types of behaviour taken as a whole, the teacher should provide a suitable environment for the student's acquisition of them.

1 3 ***Evaluation*** in education is a ***systematic process*** which enables the teacher to "measure" to what extent the student has attained the ***educational objective*** Evaluation always includes ***measurements*** (quantitative or qualitative) plus a value judgment.

1 4 To make measurements, ***measuring instruments*** must be available which satisfy ***certain requirements*** so that the results mean something to the teacher himself, the school, the student and society which, in the last analysis, has set up the educational structure

1 5 In education, measuring instruments are generally referred to as "tests"

1.6 Among the qualities of a test, whatever its nature, four are essential, namely, ***validity, reliability, objectivity*** and ***practicability.*** Others are also important, but they contribute in some degree to the qualities of validity and reliability

2 *Qualities of a measuring instrument*

The four main qualities of any measuring instrument (examination) are validity, reliability, objectivity and practicability

2 1 ***Validity:*** the extent which the test used really measures what it is intended to measure. No outside factors should be allowed to interfere with the manner in which the evaluation is carried out For instance, in measuring the ability to synthesize, other factors such as style should not compete with the element to be measured so that what is finally measured is style rather than the ability to synthesize

Validity is a concept which relates to the results obtained with a test and not to the test itself It relates more specifically to the interpretation of the results obtained by means of the test.

The notion of validity is a very relative one It implies a concept of degree, i e., one may speak of very ***valid, moderately valid*** or ***not very valid*** results

The concept of validity is ***always specific for a particular subject.*** For example, results of a test on public health administration may be of very high validity for identification of the needs of the country and of little validity for a cost/benefit or cost/efficiency analysis

Content validity is determined by the following question will this test measure, or has it measured, the matter and the behaviour that it is intended to measure?

Predictive validity is determined by questions such as the following when the results of a test are to be used for predicting the performance of a student in another domain or in another situation

To what extent do the results obtained in physiology help to predict performance in pathology?

To what extent do the results obtained during the pre-clinical years help in predicting the success of students during the clinical years?

2.2 ***Reliability*** this is the consistency with which an instrument measures a given variable.

Reliability is always connected with a particular type of consistency: the consistency of the results in time, consistency of results according to the questions, consistency of the results according to the examiners.

Reliability is a necessary but not a sufficient condition for validity. In other words, valid results are necessarily reliable, but reliable results are not necessarily valid. Consequently, results which are not very reliable affect the degree of validity. Unlike validity, reliability is a ***strictly statistical concept*** and is expressed by means of a reliability coefficient or through the ***standard error*** of the measurements made

Reliability can therefore be defined as the degree of confidence which can be placed in the results of an examination. It is the ***consistency*** with which a test gives the results expected

2.3 *Objectivity* this is the extent to which independent and competent examiners agree on what constitutes a "good" answer for each of the items of a measuring instrument.

2.4 *Practicability* depends upon the time required to construct an examination, to administer and score it and to interpret the results, and on its overall simplicity of use It should never take precedence over the ***validity*** of the test.

3. *Other qualities of a measuring instrument*

3.1 *Relevance:* this is the degree to which the criteria established for selecting questions (items) so that they conform to the aims of the measuring instrument are respected. This notion is almost identical to the one of content validity; and the two qualities are established in a similar manner.

3.2 *Equilibrium.* achievement of the correct proportion among questions allocated to each of the objectives.

3.3 *Equity:* extent to which the questions set in the examination correspond to the teaching content

3.4 *Specificity.* quality of a measuring instrument whereby an intelligent student who has not followed the teaching on the basis of which the instrument has been constructed will obtain a result equivalent to that expected by pure chance.

3.5 *Discrimination.* quality of ***each element*** of a measuring instrument which makes it possible to distinguish between good and poor students in relation to a given variable.

3.6 *Efficiency:* quality of a measuring instrument which ensures the greatest possible number of independent answers per unit of time.

3.7 *Time* it is well known that a measuring instrument will be less reliable if it leads to the introduction of non-relevant factors (guessing, taking risks or chances etc.) because the time allowed is too short.

3.8 *Length:* the reliability of a measuring instrument can be increased almost indefinitely (Spearman-Brown formula) by the addition of new questions ***equivalent*** to those constituting the original instrument

validity The extent to which the instrument measures what it is intended to measure.

reliability The consistency with which an instrument measures a given variable.

objectivity The extent to which independent and competent examiners agree on what constitutes a good answer for each of the elements of a measuring instrument.

practicability The overall simplicity of use of a test, both for test constructor and for students.

relationships between the characteristics of an examination

The diagram on the next page, suggested by G. Cormier, represents an attempt to sum up the concepts of testing worked out by a number of authors. However, no diagram can give a *perfect* representation of reality and the purpose of the following lines is to explain rather than justify the diagram.

A very good treatment of all these concepts will be found in the book by Robert Ebel entitled *Measuring Educational Achievement* (Prentice Hall, 1965)

Validity and reliability

Ebel shows that "to be valid a measuring instrument (test) *must* be both relevant and reliable." This assertion justifies the initial dichotomy of the diagram. It is, moreover, generally agreed that "a test can often, if not always, be made more valid if its reliability is increased."

Validity and relevance

According to Ebel's comments, it seems that the concept of relevance corresponds more or less to that of validity of content. In any case, both are established in a similar manner (by consensus)

By definition, a question is relevant if it adds to the validity of the instrument, and an instrument is relevant if it respects the specifications (objectives and taxonomic levels) established during its preparation

Relevance and equilibrium

It seems, moreover, that the concept of equilibrium is only a sub-category of the concept of relevance and that is why the diagram shows it as such.

Relevance and equity

It seems evident that if the instrument is constructed on the basis of a content itself determined by objectives, then it will be relevant by definition. If this is not done, then the instrument will not be relevant and consequently not valid It is equitable in the first case and non-equitable in the second However, an examination can be equitable without being relevant (or valid) when, although it corresponds well to the teaching content, the latter is not adequately derived from the objectives.

Equity, specificity and reliability

The diagram reflects the following implicit relationship· a test cannot be equitable if it is not first specific Specificity, just like equity and for similar reasons, will affect the reliability of the results.

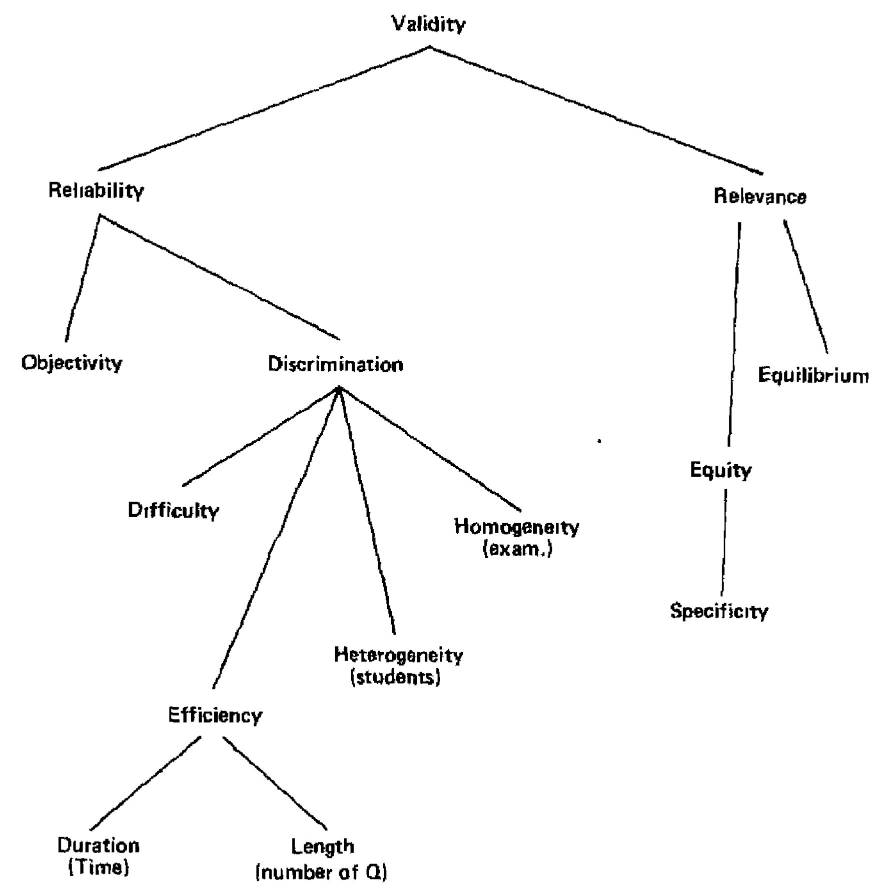
Reliability, discrimination, length, homogeneity (of questions) and heterogeneity (of students)

According to Ebel, reliability is influenced by the extent to which the questions (items) clearly distinguish competent from incompetent students, the number of items, the similarity of the items as regards their power to measure a given skill and the extent to which students are dissimilar with respect to that skill The discriminating power of a question is directly influenced (see pages 4.60 – 4.61) by its level of *difficulty*. The mean discrimination index of an instrument will also be affected by the homogeneity of the questions and the heterogeneity of the students. From the comments made above it can be seen how equity and specificity will also influence the discriminating power of the instrument

EXERCISE

Try to answer questions 22 – 25 on p. 2 47 and check your answers on p 2.48.

relationships between characteristics of an examination*



* as proposed by G. Cormier, Universite Laval, Quebec

N.B. Additional relationships to those suggested in this diagram can be established. The number of arrows has been kept to a minimum for the sake of clarity and to give a basic idea of the concept as a whole.

EXERCISE

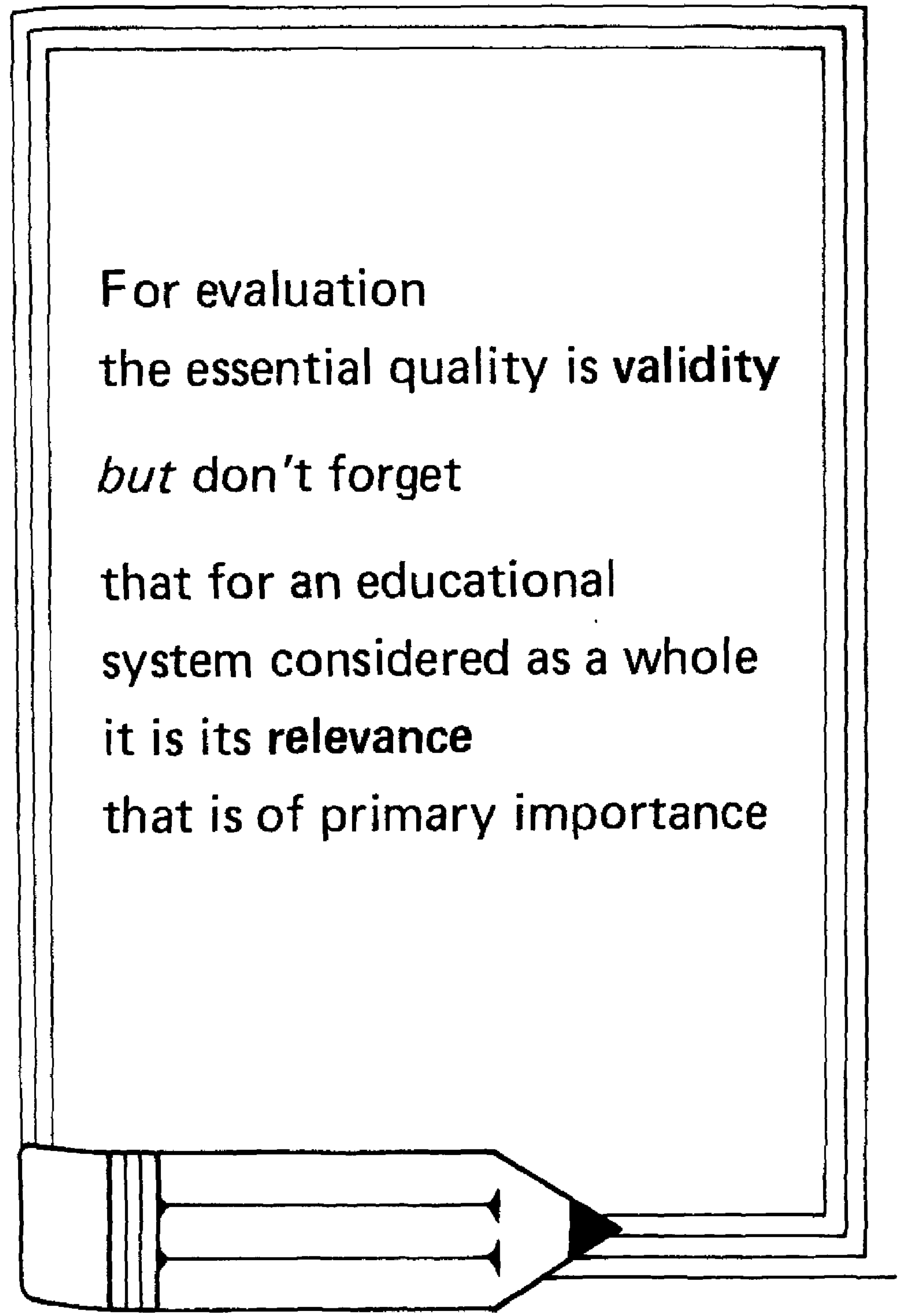
For each of the educational objectives you defined on page 1 54, describe two methods of evaluation that seem suitable to you for informing yourself and the student on the extent to which that objective has been achieved. Compare the two methods on the basis of the three criteria shown in the table below.

OBJ	Make a differential diagnosis of anaemia based on the detailed *haematological picture* described in the patient's medical record.	Validity	Objectivity	Practicability
I	Modified essay question. A series of 10 short questions based on patient's record as supplied to student (1 hour).	+++	+++	+++
II	Student given patient's record (10 mins.) followed by 15 min. oral examination.	+++	+	+

Methods of Evaluation for a class of 200 students

OBJ	The student *should be able to:*	Validity	Objectivity	Practicability
I				
II				

Check the meaning of the words validity, objectivity and practicability in the glossary, page 6.01



evaluation is a matter for teamwork

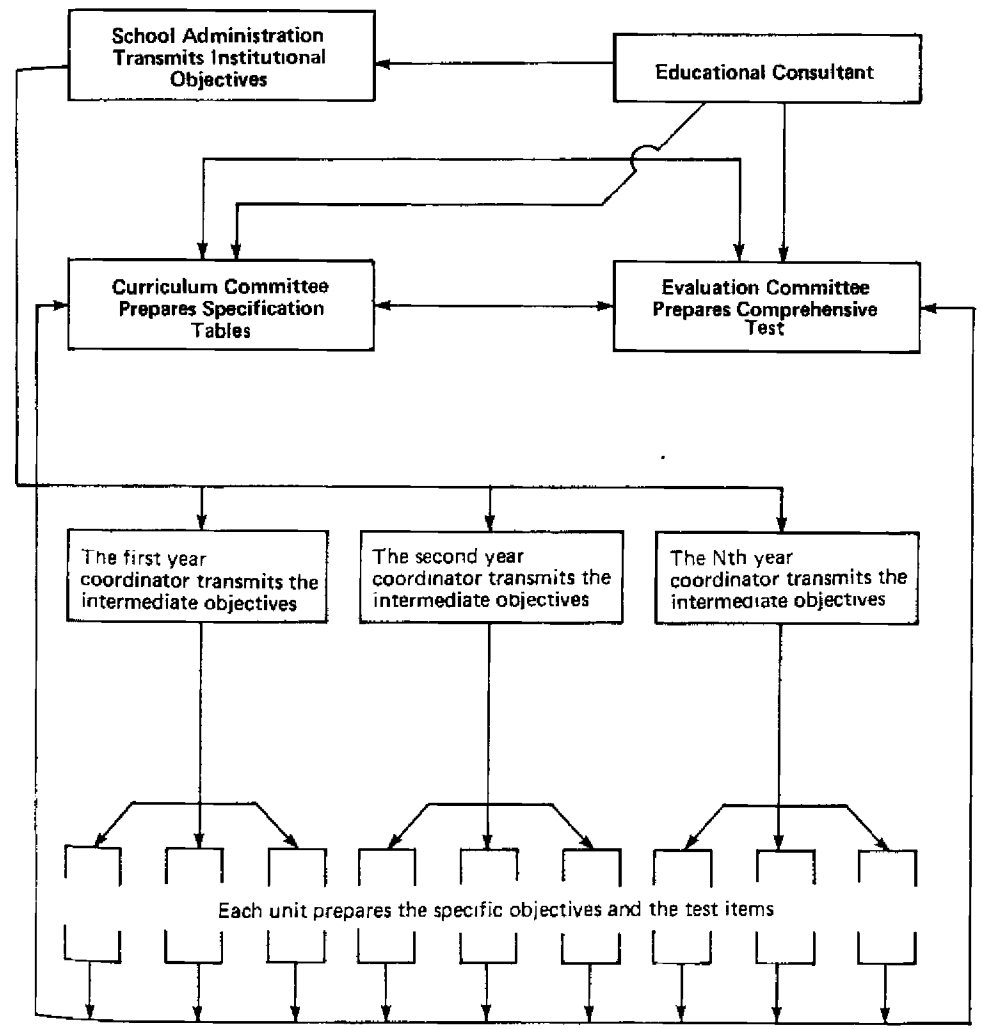
The planning of an evaluation system is obviously not simple. It is a serious matter, for the quality of health care will partly depend on it. It has been stressed many times, moreover, that it should be a group activity. We have stated in the preceding pages that evaluation must be planned jointly; that implementation of any evaluation programme is the responsibility of the teachers, in collaboration with students and the administration, that evaluation carried out jointly by a group of teachers is less likely to be erroneous than when carried out by one person; and, finally, that critical analysis of a test by colleagues is essential to its sound construction

This work performed jointly by a group of teachers calls for a coordinating mechanism The terms of reference of each group and group member must be defined explicitly and known to all. The institution's higher authorities must provide the working groups and their members with the powers corresponding to the task to be accomplished.

The diagram on the next page shows one type of organization and meets the needs of a given institution Other types of organization can be envisaged, according to existing structures and local traditions

Do not change anything that works satisfactorily . . . what is satisfactory to some, however, is not necessarily good enough for others. Teaching is a matter for teamwork.

Organizational diagram showing relationships between curriculum committee, evaluation committee and teaching units



Personal Notes

Judge the consequence of the student's not achieving the objective by answering such questions as: "If he cannot perform the objective when he leaves my instruction he is likely to . . . . .". The answer should help you decide how much energy to put into constructing a valid evaluation system to find out whether the objective is achieved as written.

R F. Mager

EXERCISE

Show graphically the type of organization (commissions, committees, administrations, etc., *with a description of their functions*) which seem desirable (in the establishment where you are teaching) for introducing (or improving) an **evaluation** system capable of providing the necessary data for checking that all **educational** objectives have been achieved (do not forget also to take account of the **decisions** you have formulated on page 2.09); then compare with page 2.41

EXERCISE

Describe the obstacles you are liable to encounter in applying the organizational plan you have imagined on the previous page and indicate tactics for overcoming each of these obstacles.

Obstacles	Tactics

EXERCISE

(Check your answers on p. 2.48)

Question 1

The main role of evaluation is

Question 2

The purpose of evaluation is to make a value judgment concerning:

A. Students and programmes

B. Students and teachers.

C Programmes and teachers.

D Students

E. None of the above.

Question 3

Thorndike's "Law of Effect" is based on the fact that

A Students learn better when they are motivated.

B. Students learn better when they play an active role.

C. Students are receptive when they understand the educational objectives which have been defined.

D. Students tend to engage in activities which have success associated with their results.

E. Students work better if the teacher makes an impression on them.

Questions 4 to 8

Indicate to which of the following each question refers.

A Formative evaluation

B. Certifying evaluation

C. Both

D. Neither

Question 4

Its main aim is to inform the student on his/her progress

Question 5

Does not preserve anonymity

Question 6

Enables the teacher to decide to replace one programme by another.

Question 7

Justifies the decision to let a student move up from the second to the third year

Question 8

Permits rank-ordering of students.

Questions 9 to 16

For each of the aims of student evaluation (list numbered 9 to 16, p. 2.18), indicate whether the appropriate measuring instrument will be of the certifying evaluation type (C) or both certifying and formative evaluation (CF)

Question 17

The four steps of the process of student evaluation are as follows

1

2.

3.

4.

Question 18

All the following steps **except one** are essential in constructing any measuring instrument.

A. Precise definition of all aspects of the type of competence to be measured.

B. Obtaining reliability and validity indices for the proposed instrument.

C. Making sure that the type of instrument chosen corresponds to the type of competence to be measured.

D. Making sure, by an explicit description of the acceptable level of performance, that the use of the measuring instrument will ensure objectivity.

E Determination of the particular behaviour expected from individuals who have or have not acquired the specified competence.

Question 19

When evaluating communication skills (domain of interpersonal relationships), all the following steps should be taken **except one:**

A. Describe specific types of behaviour showing a given affective level

B. Describe explicit types of behaviour showing the absence of a given affective level

C. Observe students in real situations enabling them to manifest the types of behaviour envisaged.

D Obtain the agreement of a group of experts on the relationship between explicit types of behaviour and the affective level envisaged.

E. Obtain the student's opinions on the way in which they would behave in specific situations

Question 20

The essential variable to be considered in evaluating the results of teaching is·

A The student's performance

B. The opinion of the teacher and his colleagues

C The opinion of the student regarding his performance.

D. The satisfaction of the teacher and the students

E. The teacher's performance.

Question 21

Which of the following is not suitable for measurement by written examinations of the "objective" type:

A. Ability to recall precise facts

B. Ability to solve problems.

C. Ability to make decisions

D. Ability to communicate with the patient.

E. Ability to interpret data.

Questions 22 and 23.

Of the following qualities can be attributed to an examination.

A = Validity B = Objectivity C = Reliability D = Specificity E = Relevance

Question 22

What quality is obtained if a group of experts agree on what constitute good answers to a test?

Question 23

What quality implies that a test consistently measures the same thing?

Question 24

The following factors, **except one,** generally affect the reliability of a test

A. Its objectivity.

B The mean discrimination index of the test questions.

C. The homogeneity of the test.

D. The relevance of the test questions.

E. The number of questions in the test.

Question 25

Which of the following test criteria is influenced by all the others?

A. Reliability.

B. Validity.

C. Objectivity.

D Specificity.

E. Relevance.

Suggested answers for the exercise on pages 2.45 – 2.47

Question	Suggested Answer	If you did not find the correct answer, consult the following pages again
1	→	2.02 – 2.06
2	E	2.08
3	D	2.14
4	A	2.15 – 2.19
5	B	
6	C	
7	B	
8	C	
9	CF	
10	CF	
11	CF	
12	C	
13	C	
14	CF	
15	CF	
16	C	
17	→	2.20
18	B	2.21
19	E	2.07 – 2.35
20	A	2.08 – 2.21
21	D	2.27 – 2.31
22	B	2.34 – 2.35
23	C	
24	D	2.34 – 2.37
25	B	