EXERCISE

You should now be able to reply "yes" to each of the following questions about the items you have just composed. Refer also to the directions given on page 4.31.

In general

1. Is the item realistic and practical?
2. Does it deal with a matter that is professionally useful and important?
3. Is it drawn up using the technical language of the profession?
4. Does it require intellectual skills of a professional kind?
5. Is it independent of every other item in the test?
6. Is it specific?
7. Does it avoid the error of giving away the correct answer by irrelevant details or extraneous data?

The essential problem

8. Is it clear?
9. Is it stated in precise terms?
10. Is it stated briefly and completely?
11. Does it contain only data related to the answer?

The distractors

12 Are they important, *plausible* answers rather than obvious distractors?

13. Do they deal with similar ideas, or data expressed in similar form?

An examination must have regard to practicability

Whether it is practicable will depend on the time necessary for its construction and administration, and the scoring and interpretation of results, as well as on the general ease of its use Examination methods that lack practicability become a heavy burden on the teacher, who will then tend to give less than due importance to the measuring instrument.

Simulation has the advantage of coming nearer to reality while permitting standardisation and protecting the patient.

the programmed examination[1]

The advantages of this relatively recent method (it appeared at the beginning of the 1960s) are so great that they should help to compensate for the difficulties attaching to its use

Briefly, its aim is to measure (by simulation on paper) the problem-solving component of clinical competence.

Like the method of multiple choice questions, it is highly objective and can be corrected by computer Its name shows that this new examination method has certain aspects identical to those of programmed teaching, where the candidate advances, step by step, through a series of consecutive clinical problems.

The method was developed in the United States of America and several types of such simulation tests can be found in the literature. They are sometimes referred to as Patient Management Problems (PMP), Clinical Simulations, etc

Objectives of the method

The aim of the method is to evaluate clinical competence:

- Ability to detect and satisfactoirly interpret abnormal signs and symptoms.
- Ability then to reach a reasonable diagnosis and to show satisfactory judgement in the choice of treatment.

Until 1963, examiners tried to find an answer to these questions by confronting the candidate with a carefully selected patient This method was effective in the past, when candidates were not very numerous. More recently faced with thousands of candidates, thousands of patients and thousands of examiners, test specialists confronted a difficulty which they rapidly recognised There were three variables: the candidate, the patient and the examiner. This represented two variables too many for a valid evaluation of the candidate.

The first research aim was to seek a valid definition of the qualities involved in what is termed clinical competence (at the level, for example, of a hospital intern) One method employed was that of the questionnaire using Flanagan's "critical incident" technique (see p. 1.09).

Through direct interviews and questionnaires, some 600 physicians were asked to describe clinical situations during which they had personally observed interns in the course of their work and had been impressed on the one hand, by examples of satisfactory clinical conduct, and, on the other by examples of unorthodox clinical conduct Three thousand situations of this type were analysed This ample documentation gave an idea of what had to be evaluated

The following step was to determine how to evaluate this "what". Numerous methods were envisaged

Silent films, in colour, of carefully selected patients were used instead of actual patients, the examiner being replaced by a series of multiple choice questions concerning the patient presented This method proved satisfactory and it is now in routine use by examining bodies

Finally another method was found (programmed testing) for evaluating the abilities of the intern when placed in a clinical situation as real as possible and called upon to face the unforeseeable problems presented by every patient.

In everyday routine the intern may be required, for example, to see a patient who has just been admitted to the medical department. He goes to the patient, gets information from him and makes a clinical examination. He must then take a certain number of decisions He calls for certain laboratory tests whose results, combined with those of the clinical examination, will lead him to reach a diagnosis and decide on a treatment The patient's condition may then improve, worsen or remain unchanged by the treatment. The situation changes, new problems appear and fresh decisions must then be taken in the light of these new data

[1] Also called "Patient Management Problems".

See also Simulation in instruction and evaluation in medicine in WHO Public Health Paper No 61, Geneva 1974

The *programmed testing* recreates, as far as possible, the changing situation represented by every patient Each patient is described in accordance with a real case history. From four to six "clinical problems" are presented following the case study with the aim of simulating a situation changing in time The patient can be followed up for several days, weeks or possibly months, just as in real life, until he is discharged either cured or with his condition improved or, if he dies, passes to the autopsy table.

At each step in time the candidate is required to make decisions; he immediately learns the results thereof and, with this fresh information, goes on to the following "choice", always concerning the same patient.

The "eraser" technique

The methodology of this type of test, as with programmed teaching, requires that the information given to the candidate is hidden from him until he has made a decision and thus becomes entitled to obtain additional information

We shall not deal at length with the different technical difficulties that had to be overcome before a satisfactory system was found. As things are at present the appropriate information ***is bidden by a completely opaque layer of ink*** which can be removed, however, with an ordinary pencil eraser, or ***revealed by a system comparable to invisible ink*** (for the formula see page 4.48).

The method can be easily used for examining a large number of candidates simultaneously.

Examples of case histories

A clinical observation, situation or study is described to the candidate and he is then asked:

1. To study the details carefully and then the list of possible decisions presented for each "choice" linked with the initial observation.

2. To choose from the list only the numbered items which seem important and appropriate

3. To erase the corresponding opaque rectangle on the answer sheet, or to discover the "consequences" by applying a special product

The candidate is reminded.

1. That except in rare instances, it is not suggested how many of the proposed decisions he should choose.

2. That information (or "consequences") will appear in the space erased for both correct and incorrect choices.

3. That since the information gradually revealed may orientate his subsequent decisions, he should consider them one after another in the order indicated.

4 But that within each "choice", the order of the numbered decisions is proposed at random although it is advisable for the candidate to re-establish a logical order in his choice

Scoring

The usual manual machine or computer method of scoring is employed, each space erased corresponding to one answer so that the candidate is unable to cancel a mistake once his choice has been made (the same applies in the case of a real patient)

The candidate is penalised whenever he makes an incorrect choice and whenever he fails to make a choice which was appropriate. The scoring is thus negative, taking into account sins of both omission and commission.

The choices proposed to candidates can be divided into three groups

(a) Appropriate; should be made with the aim of improving the patient's condition (this is indicated by the mark +1);

(b) Not indicated, should not be done and, if it is done, may be dangerous for the patient (mark −1),

(c) Neutral. of debatable importance; may or may not be done according to local conditions, teaching, customs, etc. (mark 0).

The candidate who does not make a choice regarded as suitable by the examiner or who makes a choice regarded as not indicated or dangerous is penalised.

Choice (c) has no effect on the scoring.

Consequently, this is a scoring system completely different from that for multiple choice questions where the candidate must select the best (and only) answer from several suggested.

In programmed testing he must decide to select all those choices he regards as appropriate for the treatment of the patient. He is not told how many choices he must make The same applies in medical practice, where the physician makes a choice between what should be done and what should not be done. If he is proceeding on the right lines, he makes a certain number of decisions out of all those which could be made.

Experience (immediate feedback after "erasing") gives him fresh data which will guide him towards new decisions

If he is on the wrong track, experience ("erasing") will show him his errors as they arise and give him a chance of changing his action although he will not be able to cancel out his mistakes.

Improvement of case histories

By means of test and measurement correlation studies the quality of questionnaires can be improved. The teachers who have drawn up the questions learn from the statistical study, question by question. how they can better test discriminatory qualities of judgment enabling a choice to be defined as "appropriate", "non-indicated", or "neutral"

The task is different and considerably more arduous than that involved in drawing up the usual multiple choice questions.

On the other hand, the examiners find themselves on more familiar ground and feel that they are dealing with practical clinical situations in a much more realistic way than when they had to decide on a single best choice.

The method is far from perfect and calls for constant improvement, but gives new hopes for the evaluation of the clinical competence of physicians. It makes possible the evaluation of certain qualities which were not evaluated in the past, qualities considered essential for preparing the physician to assume independent responsibility in the practice of his profession.

example of a programmed test

Specific objective: To deal in order of priority with several patients who come for treatment at the same time

Level required: To master the objectives 1, 5.1, 5.2 and 5.3 (see pp 1.29 and 1.30).

Description of the situation.

Coming into the waiting room of a children's outpatient clinic, you find 15 children accompanied by their mothers, as follows:

1. A three-year-old with a scalp affection

2. A six-month-old infant suffering from diarrhoea without outward signs of dehydration

3. A newborn infant, 10 days old, with jaundice

4. A boy, eight years old, feverish

5. A girl, three years old, with hyperthermia and dysphagia

6. A mother carrying her newborn infant under her veil

7. A 15-month-old girl with a cough and fever

8. A two-month-old infant who has suffered from diarrhoea for a week and is obviously dehydrated

9. A six-month-old infant, cyanotic, feverish, and showing signs of dyspnoea.

10. A five-year-old with expiratory bradypenia and wheezing.

11. A boy, 14 years old, with a phlegmon on his hand

12. A six-year-old girl who has suffered from abdominal pains for the last two weeks

13. An infant of seven months coming for a routine check up

14. An eight-year-old, pale but without signs of dyspnoea

15. A boy, seven years old, with arthritis of the right knee

Section A

You now decide to. (you are entitled to only one choice)

Decisions to consider	Consequences	Mark
1 Begin by examining the patients in the order of their arrival	Five minutes after beginning your examinations, the nurse calls you into the waiting room. The condition of one of the children is critical Select another decision.	- 1
2. Have measurements taken of the temperature, weight and height of all the children	Meanwhile, one of the children suffers a respiratory arrest. Select another decision.	- 1
3. Examine some of the children in priority	Select the three children that you should examine first in your consulting room. Go on to Section B and follow up these three children among the 15 on the list.	+ 1
4 Send children 7, 9 and 10 for X-ray	While on their way, one of them faints. Select another decision	- 1

Section B

Consequences	Mark
1 While you are examining this child, a child dies in the waiting room.	- 1
2. You are called into the waiting room where a child is in convulsions	- 1
3. The nurse summons you urgently.	0
4. In the waiting room, a child suffers a respiratory arrest.	-1
5 You are urgently called to the waiting room.	- 1
6 Under the mother's veil, you discover a newborn child 10 days old, cyanotic and congested T 95°F (35°C). Conjunctiva yellowish Go on to Section C.	+ 1
7. You are summoned to the waiting room.	- 1
8. W: 4.1 kg, H: 56 cm, CC: 39 cm, T: 96.4°F (35.8°C) Persistent abdominal skinfold; eyeballs sunken, cold hands and feet. Go on to Section D.	+ 1
9 W: 7.6 kg, H: 64 cm, CC 44 cm, pulse 180/min, RF: 90/min. foci of crepitant sounds in both lungs Go on to Section E.	+ 1
10. While making your auscultation you are summoned urgently.	- 1
11 During your examination, the mother of another child bursts into your consulting room with her child who is in convulsions.	- 1
12. You are urgently called into the waiting room.	- 1
13. You are urgently called into the waiting room	- 1
14. A child is in convulsions in the waiting room.	- 1
15 A child has a respiratory arrest in the waiting room.	- 1

W: Weight, H: Height; CC: Cranial circumference; T Temperature;

RF Respiratory frequency.

Section C

For this child you now decide to:

Decisions to consider	Consequences	Mark
1. Interrogate the mother	While you are doing this, the dyspnoea becomes more severe and the child becomes more congested.	- 1
2 Make a complete examination of the child.	During the examination, the child becomes cyanotic.	0
3. Request biological tests.	The moment the needle is inserted into the vein, the child has a respiratory arrest	- 1
4. Immediately treat the symptoms.	Go on to Section F.	+ 1

Section D

For this child, you now decide to:

Decisions to consider	Consequences	Mark
1 Interrogate the mother.	The child has a collapse.	- 1
2. Make a complete examination	During the examination, the child becomes cyanotic. Pulse 180	0
3. Request a blood count, sedimentation rate, urea and glycaemia	While you are inserting the needle, the child has a respiratory arrest.	- 1
4. Immediately give emergency treatment.	Go on to Section G.	+ 1

Section E

For this child, you now decide to:

Decisions to consider	Consequences	Mark
1. Interrogate the mother.	While you are doing so, the child goes into convulsions.	- 1
2 Make a complete examination.	The child's temperature rises to 106.7°F (41.5°C)	0
3 Perform a lumbar puncture	While you are doing so, the cyanosis suddenly increases.	- 1
4. Immediately give emergency treatment	Go on to Section H	+ 1

Treatment	Section F		Section G		Section H	
1 Place the child near a heat source	Temperature rises to 97 7°F (36 5°)	+1	When done Temp· 97.1°F (36.2°C)	+1	To warm it up? It already has 106 7°F (41 5°C) Treatment cancelled	-1
2 Decongest	When done, the obstruction becomes less	+ 1	Child not congested	0	Child not congested	0
3 Empty the stomach	Done	+ 1	Done	+ 1	Done	+1
4. Give oxygen	When done, the cyanosis disappears	+1	Pointless	0	When done, the cyanosis disappears	+1
5 Give a perfusion	Note under Section I the quantity of serum, the serum composition and rate of flow	+1	Note under Section J the quantity of liquid, the nature of the perfusion and the rate of flow	+1	Note under Section K the quantity of liquid, the nature of the perfusion and the rate of flow	+1
6 Give an enema of 200 cc of water at 14°C	To bring down the temperature??? Cancelled by the officer on duty	-1	Treatment unsuitable, cancelled by the resident physician	-1	Note under Section K the quantity and the nature of the liquid injected	+1
7. Aspirin	Prescription cancelled by the officer on duty	-1	Prescription cancelled by the resident physician	-1	Note under Section K the dose and the route of administration	+1
8 Cephalothin	On what basis? Prescription cancelled	-1	Not indicated. Prescription cancelled	-1	Not indicated Prescription cancelled.	-1
9 Ampicillin	There is no valid reason to give this child antibiotics	-1	Not indicated Prescription cancelled.	-1	Not indicated. Prescription cancelled.	-1
11. Ampicillin-gentamycin	You have no valid basis at this stage. Prescription cancelled for the time being	0	Not indicated. Prescription cancelled	-1	Not indicated. Prescription cancelled	-1
12. Penicillin	You have no valid basis for giving antibiotics. Prescription cancelled	-1	Prescription cancelled.	-1	Note under Section K the dose and the route of administration	+1
13. Chloramphenicol	This antibiotic is not for the newborn. Prescription cancelled.	-1	Prescription cancelled	-1	No indication Prescription cancelled.	-1
14 Hydrocortisone hemisuccinate	Not indicated Prescription cancelled.	0	Note the dose under Section J	+1	Not indicated. Prescription cancelled.	-1

Section I

Section J

Section K

Drawn up by Professor J -P. Grangaud, Test and Measurement Workshop, Timmoun, Algeria, February 1977

Note:

Special printing techniques require somewhat sophisticated and often patented apparatus A *simple* technique developed under WHO sponsorship has been published under the title ***An invisible ink process for use as an educational tool.*** in ***Information occasional publication No. 3***, 1981, B.L.A.T Centre for Health and Medical Education, Tavistock Square, London, WC1H 9JP. It is suitable for reproduction by stencil or offset processes The invisible ink and the developer can be made from products easily obtainable on the market. The simplest technique is still that described by Rimoldi* in 1955. The description of the clinical picture is typewritten on an ordinary sheet of paper; the "decisions to consider" are also typewritten, but on cards of 8 x 10 cm, and the "consequences" are on the back of the same cards When the student has made his choice, he turns over the selected card and, if it is a certifying test, the teacher notes this. The P4 packs from H. Barrows, McMaster University, Canada, are a good example of the technique.

*Rimoldi, H. J A. A technique for the study of problem solving Educational and psychological measurement, 15, 450 – 461, 1955.

EXERCISE

Try to draw up a programmed test. Take as a basis a clinical observation or an epidemiological situation. Show the result to several colleagues and ask them for constructive criticism.

List the advantages and disadvantages of this type of test.

stages of assessment

- prerequisite level testing
- pre-testing
- pre-final feedback comprehensive testing
- "subjective" impression
- final testing
- safety testing
- follow-up testing

different types of examinations during a course and their stages

Prerequisite level test

Before commencing a course it is necessary to ascertain whether the students have reached a certain level, namely the *prerequisite level* (refer back to page 1.44). A teacher must specify which knowledge he considers indispensable to ensure that the students assigned to him derive maximum benefit from the instruction he has planned for them This test shows whether all the students are at this level or whether they are not, in which case coverage of this area must be ensured by modifying the proposed instruction to bring them up to this level. If this is not done the quality of instruction must suffer. Depending on the number of students who need bringing up to this level, the teacher must decide on the type of remedy – either reference to books or additional instruction – for the students concerned, possibly with the assistance of students who have reached the level and can be given the task of "instructor". As far as possible a "repeat" for all the students should be avoided, since this would amount to ignoring the diagnosis obtained by means of the prerequisite level test.

Pre-testing

When a given course commences it is advisable to ***make sure of the level of the students with respect to the course***; on the one hand, this ***measurement of the starting level*** will permit the assessment of the real gain at the end of a course, on the other – and this has been shown experimentally – it may be found that some students are already quite advanced as regards the objectives envisaged for the course and allowance should therefore be made for this. This is a formative test (see pages 2 15 – 2.16)

Interval testing

These tests must be set *as the course proceeds* to give the student the feedback he needs in order to know where he stands after a particular period of instruction The teacher must see that these tests are, as far as possible, of the same difficulty as the final examination One way of doing this is to select at random at least three "packets" from a group of questions. These three equivalent packets (a_1, a_2, a_3, see diagram on page 4 51) will be used not only for interval testing but also for pre-testing (formative), pre-final feedback comprehensive testing (formative) and final testing (certifying) Thus, when the student reaches the final examination he will not be haunted by the idea of its difficulty, he will have been brought up to the necessary level beforehand.

Pre-final feedback comprehensive testing

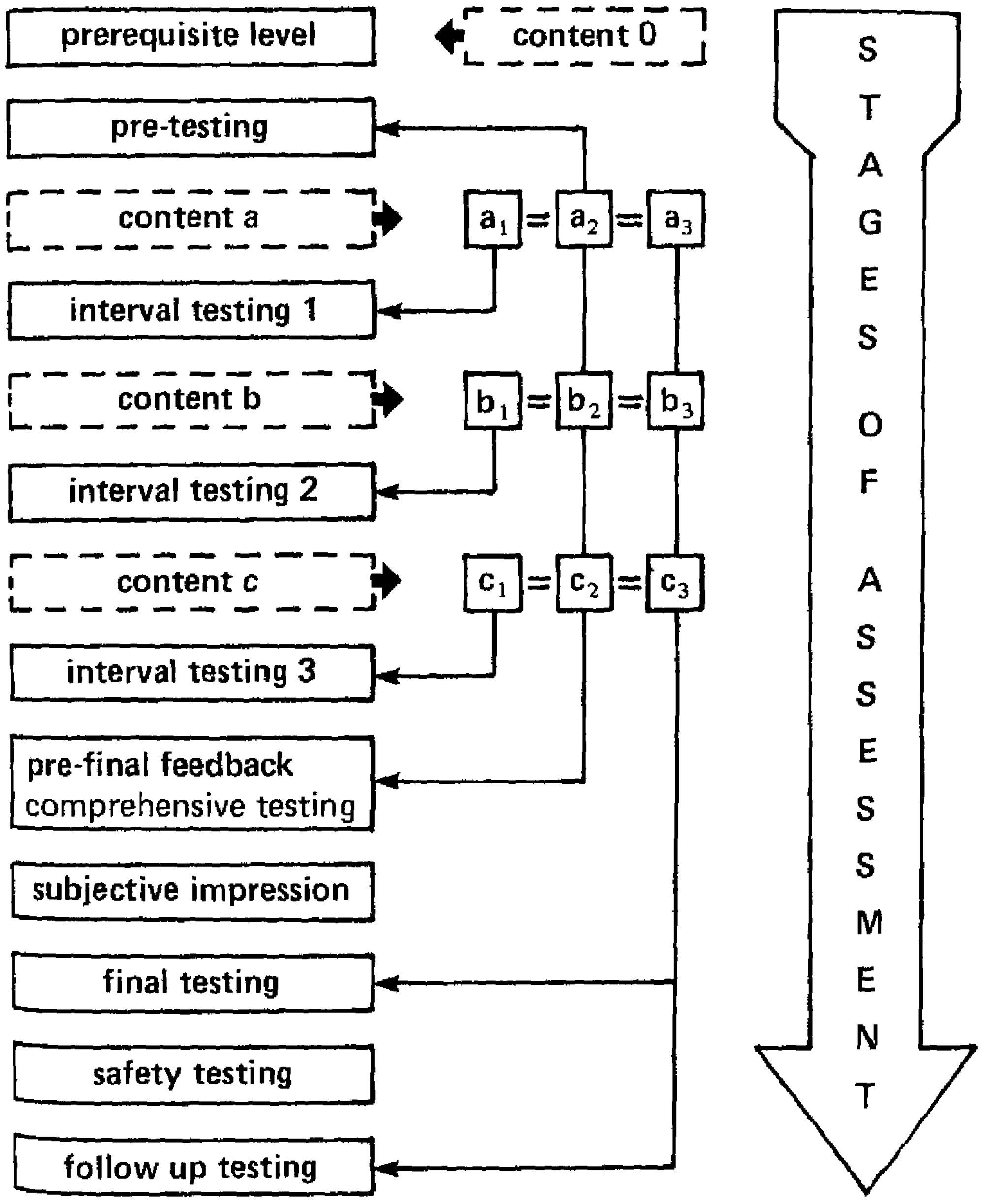
This is a test of the *formative* type set *before the final examination* (comprehensive) of a course or the year (pre-final). Its purpose is to inform the student about his level of competence (feedback) and it should not be limited to a single subject but should cover a group of subjects. This will be facilitated if the school follows an integrated curriculum. If it does not, the teacher must include questions from other fields directly relevant to understanding of the subject taught.

Subjective impression

Evaluation of this type is carried out on the basis of the teacher's *personal knowledge* of the students after contact with them during the year, he seeks to divide the students into three categories· good, average and bad. It would perhaps be preferable to divide them into two groups only: satisfactory and unsatisfactory. This evaluation should be carried out *before the final examinations* at the end of the year

Final testing

These are of different types oral, practical, traditional written, short open answer, or multiple-choice questions according to the educational objectives to be measured They are organised *after the end of a course*.



Safety testing

This should be carried out if there is an abnormal difference between the "subjective impression" and the results of the final examinations If a student who was considered satisfactory or good has a bad mark in the final examination, it is essential to re-evaluate the situation and not to give the final examination the role of final and arbitrary sanction which it has so often had in the past

Follow-up testing

This is a form of evaluation which is carried out ***sometime after completion of the course*** to determine the extent to which the student has retained the acquired level of competence.

When several measuring instruments give consistent results despite different weaknesses, the reliability of the evaluation is increased.

desirable qualities of rating scales

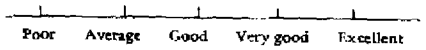
- clarity
- relevance
- precision
- variety
- objectivity
- uniqueness

factors influencing rating*

An evaluation made by a human observer is more or less objective and subject to error. The following are the factors influencing rating The list is not exhaustive!

Errors due to leniency

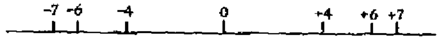
Leniency is a well known factor One means of counterbalancing this tendency is to use a scale containing only one "unfavourable" appraisal in five, for instance



In this case the appraisals will probably be distributed symmetrically around "good".

Central tendency

Examiners have a tendency not to give extreme appraisals and hence to group all candidates around the mean This ***central tendency*** may be reduced by using a scale that is wider at the centre than at the ends, for example:



The halo effect

One particular feature of a candidate sometimes seems so important to the examiner that it influences the overall evaluation Thorndike called this the ***halo effect*** However, this effect is reduced as the number of separate aspects of the problem dealt with by the evaluation is increased

The logical error

The ***logical error*** is similar to the halo effect and occurs when the examiner supposes that there is a relationship between two variables to be evaluated and that "if the first variable is of a particular order, the second will be similar". This error may be reduced if the evaluation relates to an observable element rather than to an abstraction which could lead to semantic confusion.

The contrast error

An observer who is very orderly will tend to consider, ***by contrast***, that other people are less orderly than he is, and ***vice versa***. On the other hand, people frequently believe that "others are like me" and are very surprised to see that this is not so.

The proximity error

If an observer evaluates two different factors, the evaluation of one factor tends to influence that of the other, and the shorter the interval between the two, the more pronounced the tendency (***proximity error***) will be.

*Guilford, Psychometric methods, pp 278- 280

test construction specification table

Content areas	Number of test items in relation to type of competence measured by test				
	Recall of facts	Interpretation of data	Problem Solving	n	%
Objective 1	6	3	1	10	20
Objective 2	8	2	0	10	20
Objective 3	12	6	2	20	40
Objective "n"	4	4	2	10	20
No. of items	30	15	5	50	
%	60	30	10		100

EXERCISE

Complete this specification table, making a qualitative analysis of the exercises proposed in this guide: for each exercise, decide which is the level tested – Level 1: Recall of facts; Level 2: Interpretation of data: Level 3: Problem solving; (reread page 1.39). Then calculate the percentage share of each level in the total. Check your results on the next page.

		Competence measured by the test			Number of tests		Percentage
Objective	Description of the exercise (for details, see references page 15)	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	
3	Identifying professional activities					Chapter 1	
3	Listing the main functions of a category of health personnel						
20/26	Analysis of the relevance of a programme						
5	Identifying the components of a professional task						
6	Selecting active verbs corresponding to a task						
8	Identifying the elements of a task						
8	Identifying the elements of an educational objective						
6	Drawing up specific educational objectives						
7	Drawing up contributive educational objectives						
8	Critical analysis of an educational objective						
1 to 9	Evaluation of intellectual skills with regard to educational objectives						
17	Specification of educational decisions					Chapter 2	
12	Distinguishing between formative and certifying evaluation						
13,14,16	Choice of evaluation method						

		Competence measured by the test			Number of tests		Percent-age
Objective	Description of the exercise	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	
15, 16	Comparing different methods of evaluation					Chapter 2	
17	Diagram showing an evaluation system						
18	Description of obstacles and tactics in the implementation of an evaluation system						
10 to 18	Evaluation of intellectual skills with regard to evaluation planning						
19	Description of learning situations					Chapter 3	
28	Description of the teacher's functions						
24	Selection of teaching methods						
24	Comparison between several teaching methods						
29	Construction of an organisational chart for programme implementation						
29	Description of obstacles to and tactics for setting up a new programme						
19 to 29	Evaluation of programme construction skills						
32	Listing advantages and limitations of evaluation by students					Chapter 4	
33	Construction of a practical test or a project						
33	Constructing a rating scale for attitudes						
34	Preparation of a written question						
34	Preparation of short, open answer questions						
35	Composition of multiple choice questions (MCQ)						

		Competence measured by the test			Number of tests		Percent-age
Objective	Description of the exercise	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	
36	Construction of a programmed examination					Chapter 4	
36	Completion of a specification table for an examination						
39	Calculation of the minimum pass level for MCQ tests						
40	Calculation of the difficulty and discrimination indexes of a question						
30 to 40	Evaluation of test and measurement skills						
Number of questions							
Percentage							

EXERCISE

Check your answers to the exercise on the preceding page

Your percentages should be about equal (within 10%) to those given below. Broadly, you should have found about 20% of the questions/tests at level 1 (recall of facts) and 80% at levels 2 or 3 (above level 1). At least that is what the author of this Handbook believes.

Exercises

For the exercises marked with an asterisk, you will find answer checklists in the Handbook.

See page 15 for the page numbers.

		Competence measured by the test			Number of tests		Percentage
Objective	Description of the exercise	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	
3	Identifying professional activities	–	1	2	3	Chapter 1	
3	Listing the main functions of a category of health personnel	–	–	1	1		
20 to 26	Analysis of the relevance of a programme	–	1	–	1		
5	Identifying the components of a professional task*	–	19	–	19		
6	Selecting active verbs corresponding to a task	–	1	–	1		
8	Identifying the elements of a task*	–	1	–	1		
8	Identifying the elements of an educational objective*	–	12	–	12		
6	Drawing up specific educational objectives	–	–	3	3		
7	Drawing up contributive educational objectives	–	–	2	2		
8	Critical analysis of an educational objective	–	5	–	5		
1 to 9	Evaluation of intellectual skills with regard to educational objectives*	7	13	–	20	68	31.8
17	Specification of educational decisions	–	–	1	1	Chapter 2	
12	Distinguishing between formative and certifying evaluation*	–	13	–	13		
13,14,16	Choice of evaluation method	–	5	–	5		

Exercises

For the exercises marked with an asterisk, you will find answer checklists in the handbook.

See page 15 for the page numbers.

		Competence measured by the test			Number of tests		Percentage
Objective	Description of the exercise	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	
15,16	Comparing different methods of evaluation	–	–	3	3	Chapter 2	
17	Diagram showing an evaluation system*	–	–	1	1		
18	Description of obstacles and tactics in the implementation of an evaluation system	–	–	1	1		
10 to 18	Evaluation of intellectual skills with regard to evaluation planning*	12	–	–	12	36	16.8
19	Description of learning situations	– –	–	1	1	Chapter 3	
28	Description of the teacher's functions	–	–	14	14		
24	Selection of teaching methods	–	–	12	12		
29	Construction of an organisational chart for programme implementation*	–	–	1	1		
29	Description of obstacles to and tactics for setting up a new programme	–	–	1	1		
19 to 29	Evaluation of programme construction skills*	8	11	1	20	61	28.5
32	Listing advantages and limitations of evaluation by students	–	–	1	1	Chapter 4	
33	Construction of a practical test or a project	–	–	2	2		
33	Constructing a rating scale for attitudes	–	–	1	1		
34	Preparation of a written question	–	–	1	1		
34	Preparation of short, open answer questions	–	–	6	6		
35	Composition of multiple choice questions (MCQ)	–	5	5	10		

Exercises

For the exercises marked with an asterisk, you will find answer checklists in the handbook.

See page 15 for the page numbers.

		Competence measured by the test			Number of tests		
Objective	Description of the exercise	Recall of Facts	Interpretation of Data	Problem Solving	per exercise	per chapter	Percent-age
36	Construction of a programmed examination	–	–	1	1	Chapter 4	
36	Completion of a specification table for an examination*	–	1	–	1		
39	Calculation of the minimum pass level for MCQ tests	–	1	–	1		
40	Calculation of the difficulty and discrimination indexes of a question*	–	5	–	5		
30 to 40 +	Evaluation of test and measurement skills*	12	8	–	20	49	
Number of questions		39	102	73	214	214	
Percentage		18.2 18	47.7 82	34.1			100

relative and absolute criteria tests

These two expressions are also referred to in the literature as *norm-referenced* and *criterion-referenced* tests.

It is very important to distinguish between tests based on reference to the "norm" (i.e. in accordance with the curve for the results of all the students who have taken the same test, and that is why this criterion is termed ***relative***), and tests based on reference to a "criterion" (i e., in relation to the description of an acceptable performance, that is to say, the specific educational objective fixed in advance).

An ***absolute criteria test*** is one deliberately designed to give results that can be directly interpreted in terms of the acceptable level of performance of the person tested It enables a person's performance to be evaluated in relation to a previously specified level of performance. The aim, therefore, is to determine whether a person has or has not mastered a particular task, and not to compare one person's performance with that of another or of a group of persons. A ***relative criteria test***, on the other hand, aims at enabling a valid discrimination to be made between persons on the basis of different types of performances, it is thus a competitive test.

Relative criteria tests are the ones most frequently employed for examination purposes. Unfortunately their disadvantages greatly outnumber their advantages, for if a group of students is particularly brilliant the utilisation of relative criteria tests will lead to some of them being failed although their level of performance may be satisfactory from the absolute viewpoint. On the other hand, if a given group of students has on the whole a low performance level (because the appropriate instruction has not been given, or has been poorly given, or for some other reason) the relative criteria system may allow "poor" students to pass if they are above the average of their group. The consequences may be extremely harmful for the health of the population.

If, on the other hand, a situation arises in which a certain number of persons have to be selected for admission to a given course of study, it becomes necessary to compare their performances. In that case a relative criteria test is appropriate.

If it is felt unanimously by an evaluation committee, for example, that all the students should be able to master an emergency procedure, then this can be ascertained only by an ***absolute criteria*** test. These tests are, indeed, the only ones that justify the certifying of any health worker as having demonstrated an acceptable level of performance.

It is thus theoretically possible, and even desirable, that ***all*** the students taking an absolute criteria test should "pass". That would demonstrate the high degree of effectiveness of the training programme. It would also, of course, be theoretically possible for all the students to fail.

On the other hand, a relative criteria test is, by its nature, one which will ***always*** divide the students taking it into at least two categories, those who succeed and those who do not, without any guarantee that the former are also competent.

> It basically comes down to a choice between a measurement strategy which compares people versus one that lets us know what it is that people can or cannot do.
>
> (J Popham)

A measuring technique adapted to absolute criteria tests is suggested on pages 4 62 – 4.63, namely calculation of the acceptable level of performance or minimum pass level. You will find on pages 4 65 – 4 70 the measuring techniques suitable for ***relative*** criteria tests (difficulty and discrimination indexes)

calculation of the acceptable level of performance (ALP) for a MCQ test

1 **Definition**

The acceptable level of performance is a ***threshold*** value making it possible to decide (according to ***absolute criteria***) whether a student "who knows barely enough" should be ***passed*** or ***failed***.

Calculation of the ALP for a test is not valid unless the number of MCQ is more than 30.

Use of the ALP involves an ***advance*** judgment (before the test) on the relative difficulty of each question and enables a judgment based on the test as a whole to be made

Calculation of the ALP depends on the collective decision of several teachers each of whom has first made an independent judgment.

2. **Procedure***

To calculate the acceptable level of performance (ALP) of a student for a MCQ test.

2.1 the evaluation board decides what is the correct answer to each MCQ;

2.2 the board decides which answer or answers must definitely be eliminated by the student, other than by chance;

2 3 the board calculates the ***acceptability index*** for each MCQ;

2.4 the ALP for the test as a whole is the sum of the acceptability indexes for each MCQ

The ***acceptability index*** for a MCQ is calculated as follows.

Carefully study all the choices offered (distractors) and decide which the student "who knows barely enough to pass" should be able to reject. For example, if a question offers five choices (only one of which is the correct answer) and it is deemed that the student "who knows barely enough to pass" should be able to reject one of these choices straightaway, it follows that the marginal student could obtain the correct answer by mere chance approximately one time out of four. In this case the acceptability index of the question is 0.25.

For a MCQ		with five choices	with four choices
– if all the distractors are equivalent, the index	=	1/5 = 0 20	1/4 = 0 25
– if one distractor must be eliminated, the index	=	1/4 = 0.25	1/3 = 0.33
– if two distractors must be eliminated, the index	=	1/3 = 0.33	1/2 = 0 50
– if three distractors must be eliminated, the index	=	1/2 = 0 50	1/1 = 1.00
– if four distractors must be eliminated, the index	=	1/1 = 1.00	

Let us take two MCQ which are worded identically but where the choice of answers is different.

Question:

Which of the following values corresponds to the number of red cells per mm^3 of blood in a healthy adult?

	1		2
A	500 000	A	4 000 000
B	1 000 000	B	4 500 000
C	2 000 000	C	4 750 000
D	3 000 000	D	5 000 000
E	5 000 000	E	5 250 000

In case 1, an acceptability index of 1.00 could be considered while in case 2 it could be 0.25

*This procedure was developed from an article by L. Nedelsky (Absolute grading standards for objective tests Educ Psycholog Meas., 14, 3–19 1954).

3. **Comments**

The ALP has little value if it is not based on a detailed analysis of each of the questions in a test, including consideration of incorrect choices just as much as of correct answers.

The validity of the estimate of the ALP also depends on obtaining ***independent*** judgments from several teachers who have paid attention to the educational objectives and the level for which the examination is intended The usefulness of the estimate will be the greater the larger the number of teachers involved

When the differences between the judgements obtained are relatively small, the extremes can serve to define a "grey zone" below which the results will be regarded as distinctly inadequate (failure) and above which the results will clearly indicate a success. For example, if the ***mean*** of the estimates of one teacher for the ALP of a test is 43% whereas two other teachers obtain figures of 45% and 47%, respectively, then it could be recommended that any score below 43% should be regarded as a failure, that any score above 47% be regarded as a success, while a score between 43% and 47% should be regarded as being in a grey zone It would remain to be defined what should be done in the latter case

If the differences between the judgments obtained by several teachers are large then the criteria of the educational objectives should be revised.

EXERCISE

Now . . . . . calculate the acceptable pass level for all the MCQ you drew up on page 4.39.

1. Award of a score to each student

A practical, simple and rapid method is to perforate on *your* answer sheet the boxes corresponding to the correct answer. By placing the perforated sheet on the student's answer sheet the raw score (number of correct answers) can be found almost automatically

	A	B	C	D	E		A	B	C	D	E
1.	□	□	☒	□	□	11	☒	□	□	□	□
2.	□	□	□	□	☒	12	□	□	☒	□	□
3.	□	□	☒	□	□	13	□	□	☒	□	□
4.	□	□	□	☒	□	14	☒	□	□	□	□
5.	□	☒	□	□	□	15.	□	□	□	☒	□
6.	□	□	☒	□	□	16.	☒	□	□	□	□
7.	☒	□	□	□	□	17.	□	☒	□	□	□
8.	□	□	□	☒	□	18.	☒	□	□	□	□
9.	☒	□	□	□	□	19.	□	□	☒	□	□
10.	□	☒	□	□	□	20.	□	☒	□	□	□

steps in item analysis (relative criteria tests)

1. award of a **score** to each student

2. **ranking** in order of merit

3. identification of **groups**: high and low

4. calculation of the **difficulty index** of a question

5. calculation of the **discrimination index** of a question

6. critical **evaluation** of each question enabling a given question to be retained, revised or rejected

1. Award of a score to each student

A practical, simple and rapid method is to perforate on *your* answer sheet the boxes corresponding to the correct answer. By placing the perforated sheet on the student's answer sheet the raw score (number of correct answers) can be found almost automatically

	A	B	C	D	E		A	B	C	D	E
1.	□	□	☒	□	□	11	☒	□	□	□	□
2.	□	□	□	□	☒	12	□	□	☒	□	□
3.	□	□	☒	□	□	13	□	□	☒	□	□
4.	□	□	□	☒	□	14	☒	□	□	□	□
5.	□	☒	□	□	□	15.	□	□	□	☒	□
6.	□	□	☒	□	□	16.	☒	□	□	□	□
7.	☒	□	□	□	□	17.	□	☒	□	□	□
8.	□	□	□	☒	□	18.	☒	□	□	□	□
9.	☒	□	□	□	□	19.	□	□	☒	□	□
10.	□	☒	□	□	□	20.	□	☒	□	□	□

2. Ranking in order of merit

Assuming that the scores of 21 students have been obtained (alphabetical list on the left), this step consists merely in ranking (listing) students in order of merit (in relation to the score) proceeding from the highest to the lowest score Let us assume the list as under A and then rank the students to obtain distribution B, ranging from 4 to 27.

Student	Score
Albert	7
Alfred	13
Andrew	19
Ann	25
Brian	27
Christine	19
Elizabeth	17
Emily	24
Felicity	16
Frances	14
Frank	26
Fred	17
Harriet	11
Ian	17
John	14
Jennifer	21
Margaret	16
Michael	9
Paul	16
Peter	4
Philip	16

Order	Student	Score
1	Brian	27
2	Frank	26
3	Ann	25
4	Emily	24
5	Jennifer	21
6	Christine	19
7	Andrew	19
8	Elizabeth	17
9	Ian	17
10	Fred	17
11	Felicity	16
12	Margaret	16
13	Paul	16
14	Philip	16
15	Frances	14
16	John	14
17	Alfred	13
18	Harriet	11
19	Michael	9
20	Albert	7
21	Peter	4

3. Identification of high and low groups

Ebel[1] suggests the formation of "high" and "low" groups comprising only the first 27% (high group) and the last 27% (low group) of all the students ranked in order of merit.

Why 27%? Because 27% gives the best compromise between two desirable but contradictory aims:

1. making both groups as large as possible;
2. making the two groups as different as possible.

Truman Kelley showed in 1939 that when each group consists of 27% of the total it can be said with the highest degree of certainty that those in the high group are really superior (with respect to the quality measured by the test) to those in the low group. If a figure of 10% were taken, the difference between the two means of the competence of the two groups would be greater but the groups would be much smaller and there would be less certainty regarding their mean level of performance.

Similarly, if a figure of 50% was taken the two groups would be of maximum size but since the basis of our ranking is not absolutely accurate, certain students in the high group would really belong to the low group, and *vice versa*.

While the choice of 27% is the best, it is, however, not really preferable to 25% or 33%; and if it is preferred to work with $\frac{1}{4}$ or $\frac{1}{3}$ rather than with the somewhat odd figure of 27% there is no great disadvantage in so doing.

For the rest of our analysis we shall use 33%.

4. Calculation of the difficulty index of a question

Difficulty index

Index for measuring the easiness or difficulty of a test question. It is the percentage (%) of students who have correctly answered a test question, it would be more logical to call it the easiness index. It can vary from 0 to 100%

Calculation

The following formula is used:

$$\text{Difficulty index} = \frac{H + L}{N} \times 100$$

where H = number of correct answers in the high group

L = number of correct answers in the low group

N = total number of students in both groups

(Do exercise on page 4.71).

5. Calculation of the discrimination index of a question

Discrimination index

An indicator showing how significantly a question discriminates between "high" and "low" students. It varies from −1 to +1

Calculation

The following formula is used:

$$\text{Discrimination index} = 2 \times \frac{(H - L)}{N}$$

(Do exercise on page 4.71).

[1] Ebel, R L, (1965) Measuring educational achievement, Prentice Hall, pp 348 – 349

6 Critical evaluation of a question

This is based on the indexes obtained.

Difficulty index: the ***higher*** this index the ***easier*** the question; it is thus an illogical term It is sometimes called "easiness index", but in the American literature it is always called "difficulty index".

In principle, a question with a difficulty index lying between 30% and 70%* is acceptable (in that range, the discrimination index is more likely to be high).

If for a test you use a group of questions with indexes in the range 30% – 70%, then the mean index will be around 50%. It has been shown that a test with a difficulty index in the range of 50% – 60% is very likely to be reliable as regards its internal consistency or homogeneity.

Discrimination index: the ***higher*** the index the more a question will distinguish (for a given group of students) between "high" and "low" students. When a test is composed of questions with high discrimination indexes, it ensures a ranking that clearly discriminates between the students according to their level of performance, i.e , it gives no advantage to the low group over the high group In other words, it helps you to ***find out who are the best students***

It is most useful in preparing your question bank. Using the index**, you can judge questions as follows.

0.35 and over		Excellent question
0 25 to 0 34	:	Good question
0.15 to 0 24	:	Marginal question – revise
under 0.15	:	Poor question – most likely discard

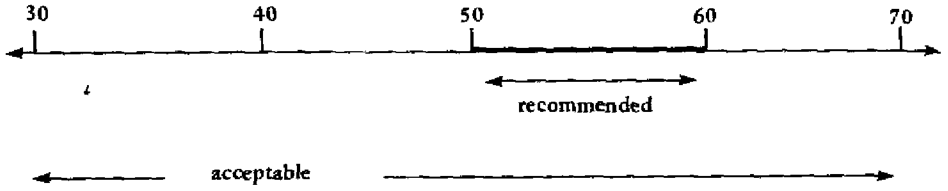
*Some authors give values between 35% and 85%.

** Remember that the index has an indicative rather than an absolute value

uses of indices
aim: review of questions

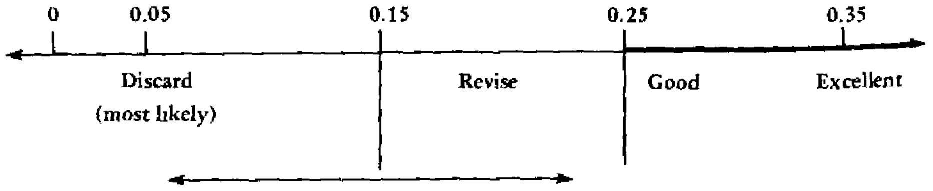
Difficulty Index

$$\text{diff. ind.} = \frac{H + L}{N} \times 100$$



Discrimination Index

$$\text{dis. ind.} = \frac{2\,(H - L)}{N}$$



EXERCISE

Given a group of 21 students (see page 4.67) Using 33% of them to constitute a high group of 7 and a low group of 7 (33% of 21), the following table shows the answers given by those two groups (high and low) to 10 multiple choice questions (numbered from 1 to 10 in the first column). The correct answer for each of those ten questions is given correspondingly in the second column. In the 14 consecutive columns are shown the answers given by each student *to each question.*

Question No.	Correct Answer	Ranking in order of merit														H	L	H + L	H - L	DIF. IND.	DIS. IND.
		Brian	Frank	Ann	Emily	Jennifer	Christine	Andrew	Frances	John	Alfred	Harriet	Michael	Albert	Peter						
1	B	B	B	B	B	B	B	B	B	B	B	B	E	E	E	7	4	11	3		
2	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	7	7	14	0		
3	A	B	A	B	A	B	B	A	B	B	A	B	A	A	A	3	4	7	-1		
4	E	E	E	E	C	E	E	C	E	C	C	A	E	-	C	5	2	7	3		
5	B	B	C	B	B	C	E	C	C	C	E	E	E	E	E	3	0	3	3		
6	D	D	D	D	C	D	D	D	E	D	E	E	D	E	E					57	0.57
7	A	C	C	C	C	C	C	C	C	C	C	C	C	C	C					0	0
8	C	B	B	B	C	B	C	B	B	B	B	B	B	C	C					28	0
9	E	E	E	E	E	E	E	E	C	E	E	E	C	B	A					71	0.57
10	C	C	C	C	C	C	A	C	C	B	-	C	D	B	A					57	0.62
		33%							33%												
		High group							Low group												

Now try to:

1. Calculate H - L for questions 6 to 10.
2. Calculate the *difficulty index* and the *discrimination index* for questions 1 to 5.

Check your results on the next page.

Check your answers

Question No.	Correct Answers	Ranking in order of merit														H	L	H + L	H - L	DIF. IND.	DIS. IND.
		Brian	Frank	Ann	Emily	Jennifer	Christine	Andrew	Frances	John	Alfred	Harriet	Michael	Albert	Peter						
1	B	B	B	B	B	B	B	B	B	B	B	B	E	E	E	7	4	11	3	78	0.42
2	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	7	7	14	0	100	0
3	A	B	A	B	A	B	B	A	B	B	A	B	A	A	A	3	4	7	-1	50	-0.14
4	E	E	E	E	C	E	E	C	E	C	C	A	E	-	C	5	2	7	3	50	0.43
5	B	B	C	B	B	C	E	C	C	C	E	E	E	E	E	3	0	3	3	21	0.42
6	D	D	D	D	C	D	D	D	E	D	E	E	D	E	E	6	2	8	4	57	0.57
7	A	C	C	C	C	C	C	C	C	C	C	C	C	C	C	0	0	0	0	0	0
8*	C	B	B	B	C	B	C	B	B	B	B	B	B	C	C	2	2	4	0	28	0
9	E	E	E	E	E	E	E	E	C	E	E	E	C	B	A	7	3	10	4	71	0.57
10	C	C	C	C	C	C	A	C	C	B	-	C	D	B	A	6	2	8	4	57	0.62
		33%							33%												

Conditions for the application of this procedure for item analyses, in particular:

1. it applies to *relative* criteria tests (the procedure leads to a choice of questions that tend to maximise variance and ensure discriminatory ranking),
2. it is applicable only to questions scored dichotomously (1;0),
3. it should not be applied if the total number of students is very small (a minimum of 20 students could be proposed as a "pragmatic" criterion).

Question analysis card

To facilitate the construction of a *question bank*, it is advisable to enter the statistical results for each question on a separate card.

These cards as a whole will constitute the "bank"

The front and the back of the card of this type could be as follows

Front

Course	Date	Nature of test	Group	Size of group	Chosen answers						Difficulty index	Discrimination index	Comments
					A	B	C	D	E	Blank			
2nd year	6.72	MCQ	High	60	55	2	3	0	0	0	60	0.63	
			Low	60	17	5	3	7	28	0			
Clin Med	6 72	MCQ	High	10	7	0	2	0	0	1	57	0 30	
			Low	10	4	0	1	1	4	0			
2nd year	6.73	MCQ	High	62	56	2	3	1	0	0	60	0.61	
			Low	62	18	8	3	8	24	1			
2nd year	6.74	MCQ	High	70	60	0	2	3	5	0	59	0 57	
			Low	70	20	7	4	2	30	6			
Speciality Board	7.74	MCQ	High	20	18	0	2	0	0	0	80	0.20	
			Low	20	14	0	0	1	5	0			

Back

Subject	Endocrine System	Nature of question MCQ
Objective tested	Ability to explain the physiological functioning of the thyroid gland	
Domain	Intellectual skills – Level 1 (recall of facts)	
Question	Which of the following produces an increased secretion of thyroid hormone in a normal subject?	
Answers	A Administration of TSH B. Administration of thiocyanate C. Administration of propylthiouracil D. Administration of thyroxine E. Some other treatment	
Reference	Sternberg, Chapter 2, page 112 – prepared by Mr X in February 1972	

Personal Notes

EXERCISE

Check your results on page 4.82.

Question 1

The administration of a test before the beginning of a learning period (formative pre-testing) has the following advantages *except one*:

A. To modify educational objectives of that period.

B. To provide ways for less well prepared students to catch up.

C. To modify the required pass level (mark).

D To provide a base from which to measure real progress.

E. To exclude weak students from the learning period.

Question 2

All the following stages, *except one*, are recommended for scoring tests of the "essay" type:

A. Write the elements of the answer for each of the questions asked.

B. Correct the answers question by question rather than student by student.

C. Determine the pass score on the basis of a sample of answers.

D. Correct the answers while preserving the anonymity of the students.

E. Identify three levels only: honour, pass, fail.

Question 3

The *content validity* of a written test is *usually* obtained by means of:

A. Collective and careful review of the questions.

B. Pearson's correlation coefficient.

C. Factor analysis.

D. An "inter-rater" reliability coefficient.

E. A mean discrimination index.

Questions 4 to 6

A test with 50 questions is administered to a group of 45 students. There is a choice of five answers to every question. Only one of these choices is the correct answer. One point per correct answer is allocated in calculating the total score.

Question 4

Assuming that none of the students have any knowledge of the test subject (i.e. they choose their answers by guessing), which of the following will be closest to the mean score of the group?

A. 0

B. 5

C. 10

D. 15

E. 25

Question 5

On dividing this group of 45 students into 3 groups of 15 each, on the basis of the total score of each student, it is found that, for the first question, nine students out of 15 in the high group and three out of 15 in the low group have given the right answer. For this question the *difficulty index* is:

A. 12%

B. 27%

C. 30%

D. 40%

E. 60%

Question 6

Under the same conditions, the *discrimination index* is:

A. 0.12

B. 0.27

C. 0.30

D. 0.40

E. 0.60

Question 7

On the basis of these indexes, which of the following decisions would you take concerning this question?

A. It should be discarded from the question bank.

B. It should be referred to a drafting committee for revision.

C. It should be retained in the bank as it is.

D. A decision other than A, B or C.

Questions 8 and 9

The following data concern a multiple choice question set to 300 students, the correct answer being D.

	Choice of answers					
	A	B	C	D	E	No answer
High group (100)	22	1	10	67	0	0
Low group (100)	46	5	16	33	0	0

Question 8

These data show that:

A. half the students answered the question correctly,

B. all the distractors were of good quality;

C. the question was of high validity;

D. the question was not very relevant.

Question 9

In view of these data, the examination board may decide:

A. that this question should be reviewed since it is insufficiently discriminatory;

B. that this question should be discarded from the question bank;

C. that this question is of low validity;

D. none of the above.

Question 10

What could *generally* be expected on doubling the length of a test whose mean discrimination index is 0.52 (by adding questions more or less equivalent to the previous ones)?

A. A certain increase in the reliability and the validity of the test

B. Only a certain increase in the reliability of the test

C. Only a certain increase in the validity of the test.

D A certain decrease in the reliability and validity of the test.

E. No effect on either the reliability or the validity of the test.

Questions 11 to 16

Use the following key in answering this series of six matching type questions·

A = traditional oral test

B = written test of the essay type

C = so-called written "objective" test (MCQ)

D = standardised practical test, or written and oral simulation tests (programmed examination)

Indicate the type of test *most suitable* for evaluating each of the following performances·

Question 11

Recall of concepts.

Question 12

Ability to solve problems.

Question 13

Ability to communicate satisfactorily with the patient.

Question 14

Verbal expression.

Question 15

Skill in examining the patient

Question 16

Ability to make a synthesis.

Question 17

The system of "relative" criteria of competence implies the following consequences, *except one*. Which?

A Leads to an embarrassing disagreement among those responsible for applying the resultant decisions.

B. Leads to the failure of certain students in a particularly competent group

C. Enables one group to become the arbiter of the standards according to which it is judged.

D. Enables "low group" students, who are however superior to the mean of the whole group to which they belong, to pass.

E. Creates an arbitrary fluctuation in the desirable level of competence at a given moment

Questions 18 and 19

The author of the following multiple choice question was asked to establish its acceptability index

> The diameter of a normal erythrocyte (according to Wintrobe) expressed in μm (microns) is equal to:
>
> A. 4.5 C. 7.5 D. 8.5
>
> B. 6.5 E. 10.5

He felt that a student who "knew just enough to pass" should be able to reject right away choices A and E.

Question 18

Indicate which among the following values of the acceptability index corresponds to the author's choice:

A. 0.10 C. 0.25 D. 0 33

B 0.20 E. 0.50

Question 19

If the item C was not included, what then would be the acceptability index?

Question 20

According to the theories about absolute or relative criteria tests, all the following statements are correct *except one.* Indicate which is false:

A. The calculation of the discrimination index provides a statistical datum applicable to absolute criteria tests

B. The calculation of the acceptable level of performance (ALP) of a test is applicable to criterion-referenced tests.

C. The acceptable level of performance (ALP) of a test is equal to the sum of the acceptability indexes of each question.

D. The value of the difficulty index influences the value of the discrimination index

Performance assessments designed to measure competence for a job or task are inescapably imperfect because of measurement errors, and because task components can never represent the total job.

If you find this Handbook too complicated you may wish to refer to another publication of the World Health Organization: *Teaching for better learning: A guide for teachers of primary health care staff* by F. R. Abbat, WHO, Geneva, 1980.

please

. . . . . if you know how to define specific objectives

. . . . . if your colleagues turn green with envy on reading your criteria

. . . . . if you are able to choose the most suitable teaching technique

. . . . . if you can put it into practice

. . . . . if you are a leader

. . . . . if your students admire you

. . . . . if your examinations are valid

. . . . . if your scores are objective

do not be influenced by all this . . !

. . . and over-estimate the importance of your own subject.

do not forget relevance . . !

. . . . . the relationship between your teaching and the institutional objectives derived from community health needs.

Answers suggested for the exercise on pages 4.75 – 4.80.

Question	Suggested answer	If you did not find the correct answer, consult the following pages again
1	E	2.15 and 2.16, 4.49 – 4.52
2	C	4.28
3	A	2.33
4	C	simple "rule of 3"
5	D	4.65 – 4.72
6	D	
7	C	
8	A	
9	D	
10	A	2.36 and 2.37
11	C	2.22, 2.30 and 2.31, 4.22 – 4.40
12	D	
13	D	
14	A	
15	D	
16	B	
17	A	4.61 and 4.62
18	D	4.62 and 4.63
19	E	
20	A	4.61 – 4.63