DEVELOPING CLASSROOM TESTS

By:
Dr. Haytham Ahmed Zakai
Medical Laboratory Technology Department
Faculty Applied Medical Sciences

Introduction.

It would be difficult to consider classroom activities for very long before running into the nemesis of students: tests. They are everywhere. Tests to get into schools; tests to be tracked by; tests for achievement; and tests to get out of school.
It seems imperative that the ordinary classroom teacher be prepared to develop good tests. Unfortunately, while many teachers devise tests and give and grade them, few teachers have the sophistication needed to develop even reasonably adequate tests. Moreover, some teachers are unaware of the purposes that an ordinary classroom test can and often does serve. Few teachers realize what a powerful tool tests can be. In their hands, tests used properly can bring out the desired achievement to students. Sadly enough, however, tests used improperly can have devastating effects on student morale, self-concept, attitudes toward school, subject matter and toward the teacher.

To test or not to test.

Evaluation of student progress is a major aspect of the teacher's job. One tool that the teacher needs in his/her evaluative repertoire is a test to provide specific information relevant to the judgment or evaluation must be made about the students.

What functions do tests serve?

Diagnostic function. Pretests indicate how prepared students are to profit from new instruction and to what degree they have mastered previous concepts and skills. Also tests serve a diagnostic role by highlighting students' strengths and weaknesses as well as trouble spots needing special remedial attention.

Motivational function. Tests motivate students to study the material assigned. They want to succeed academically, they fear failure, or they want to compete. Also, when given prompt feedback, students are motivated to improve, to alter their mistaken concepts, and in some instances to delve more deeply into the subject matter.

Self-evaluative function. When students get feedback on tests, they learn how others appraise their efforts and abilities. In turn, they develop their own self-evaluation skills.

Instructional function. Tests cause students to review material and to consolidate and integrate ideas. In addition, preparing for tests provides them with an overall review of the lesson and its important aspects.

Improving teacher effectiveness. Tests indicate to the teacher how well the material was covered, what areas need improvement, and how he/she might better organize the material for a clearer presentation to the students.

Few notes about objectives.

Cognitive domain include:

Knowledge. Involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure or setting. To use an analogy, if one thinks of the mind as a file, the problem in knowledge test situation is that of finding in the problem or task the appropriate signals, cues, and clues which will most effectively bring out whatever is filed or stored.

Comprehension. This represents the lowest level of understanding. It refers to a type of understanding or apprehension such that the individual knows what is being communicated without relating it to the other material or seeing its fullest implication.

Application. The use of abstractions in particular and concrete situations. The abstractions may be in the form of general ideas, rules of procedures, or generalized methods. The abstractions may also be technical principles, ideas, and theories which must be remembered and applied.

Analysis. The breakdown of communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between ideas expressed are made explicit. Such analyses are intended to clarify the communication, to indicate how the communication is organized, and the way in which it manages to convey its effects, as well as its basis and arrangements.

Synthesis. The putting together of elements or parts so as to form a whole. This involves the process of working with pieces, parts, elements, etc., and arranging and combining them in such a way as to constitute a pattern or structure not clearly there before.

Evaluation. Judgment about the value of material and methods for given purposes. Quantitative and qualitative judgments about the extent to which material and methods satisfy criteria. Use of a standard of appraisal. The criteria may be those determined by the student or those, which are given to him.

Planning Tests

It is essential that teacher-made tests be well thought out in advance, accurately reflect the instructional emphasis, and clearly coincide with the intended terminal student behaviors (objectives). Assuming that you have written meaningful and appropriate behavioral objectives, your students will be keyed in to the important aspects and processes that will be covered in your instruction. Moreover, the objectives serve as guides to students, telling them what to expect on the test and as guards for the teacher so that only questions related to his/her teaching are included on the test.

How can you avoid developing tests that might do more harm than good? Well, if you have taken the time to write behavioral objectives for a unit, in part you have already planed what you should test the students over. What you might need is an organizational tool called a "*Specification Chart*" to help ensure that you use them. Specification charts lay out your objectives on a two-way grid. Along one axis you list a brief outline of the content you consider essential and relevant to the unit you are teaching. Along the other axis, you specify the levels of cognitive functioning that you are interested in developing in your students.

When developing test, look at the specification chart. It specifies exactly what content should be tested and, moreover, at precisely what levels the question should be asked. One of the advantages of specification charts is that the teacher's task is made easier at test times. Also, the students are not tested over material or skills for which they received no instruction. In addition to that, the teacher is alerted to a potential overemphasis of minor points and possible inadequate coverage of major concepts or important skills.
Last, and perhaps most importantly, the use of specification chart can help increase the validity (content validity) of the teacher made test.

Besides ensuring agreement between test items and objectives, the specification chart can further increase the validity of the test being developed if for each row on the chart the relative importance of that section of the content outline is specified. The teacher might indicate the relative importance by assigning weights or percentages. The relative weight for a unit is:

$$\frac{\text{Class time spent on a particular Section of the unit}}{\text{Total class time allocated to the whole unit}} = \text{Relative Weight or \%}$$

In addition to these primary weights assigned to each row of content, the teacher must also indicate a secondary weight for each objective within a row.

The final aspect to be considered in planning a test is the number of items to be used. Obviously, this is partly a matter of individual preference and is dependent on several

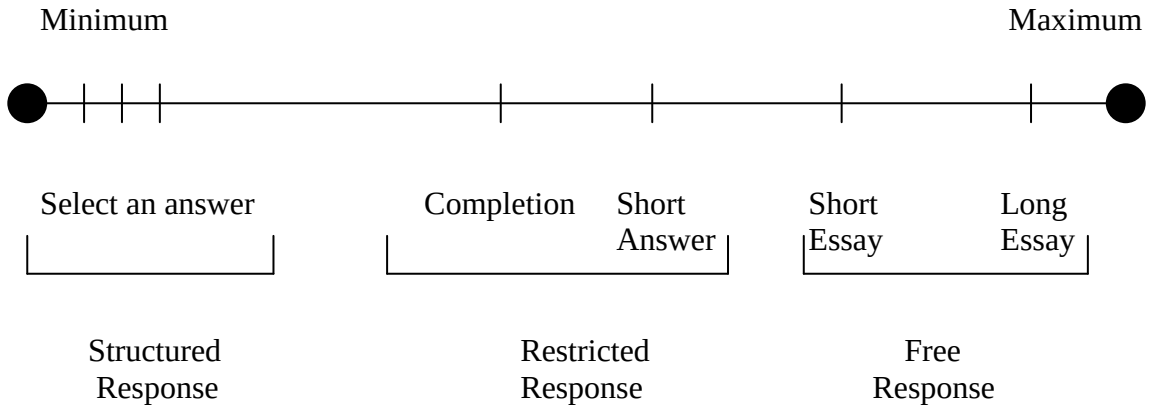
considerations: type of test items, amount of subject matter covered, ages and educational level of students, degree of reliability desired, etc. However, in the ordinary classroom the advantage of smaller, more frequent tests seems to outweigh any advantages associated with fewer but longer tests.

Using the primary and secondary weights from the specification chart, it is relatively simple to calculate how many test questions should be constructed for each objective.

Developing Test Items

Choosing items

One of the most important elements in planning your test concerns the type of test questions you decide to use. The possible choices open to you can be arranged along a continuum. At one end of the continuum, students have a minimum degree of freedom when responding to select-a-response question. The students latitude in generating an answer gradually increases as you move from left to right along the continuum. The maximum degree of latitude occurs when the student writes an extended essay response.



Advantages and limitations regarding construction, scoring, and sampling of free-response, restricted-response, and structured-response are shown in the following table.

		Free-response	Restricted-response	Structured-response
		Long essay & short Essay	Completion & short answer	MCQ, T&F, and Matching
CONSTRUCTION	Time	Can be constructed relatively quickly	Somewhat more time is required for constructing completion items than is needed for essay items	These items require the most time for construction, especially MCQ.
	Ease	Relatively easy to develop the few items needed for a test of this type	More difficult to construct than essay items because of the need to delimit the problem	Difficult to construct T&F items that are not obvious or picayune, distractors in MCQ are sometimes difficult to develop
	Behavior	Aptly suited to testing higher order cognitive behaviors. Not efficient in measuring merely knowledge level. Only for suited to measure creative abilities and organizational skills	Primarily restricted to measuring knowledge and comprehension level objectives because of limited freedom in responding	MCQ are suitable for testing most higher order skills whereas T&F and matching items assess only knowledge level behaviors
SCORING	Time	Very time consuming to grade accurately	Can be scored relatively quickly in comparison to essay items	Scoring takes very little time to complete
	Ease	Difficult to score. Require a competent authority to make evaluative judgment	Easier to score because of the short answer, but still require a competent judge	Scoring is easy and simple
	Behavior	Subject to high degree of scorer unreliability due to the halo effect, extraneous factors, and an indefinite scoring key	More objectively scored but unreliable to the extent that unintended answers must be evaluated by the teacher	High degree of reliability in scoring is possible since student responses are usually limited to a single letter or number
SAMPLING		Only a limited amount of material can be sampled in each test. Content validity is usually lacking when there are many objectives	While adequate sampling of material is possible, objectives requiring higher order behaviors are not adequately sampled with a consequent lack of content validity	Highly representative samples of objectives and content are possible with MCQ. T&F and matching items generally lack content validity when higher order skills are taught
MISC		Students produce their own response. They are encouraged to study relationships and important concepts as opposed to mere facts. Wild guessing is not a problem but padding frequently occurs.	Students tend to study only isolated facts. Students must recall information rather than merely recognize the correct answer.	Student might be inclined to study only facts for T&F and matching items. Cost to produce is greater than with other forms.

Writing Items

There are some general rules to consider that apply to all types of teacher-made tests. These guidelines are a matter of common sense since they help to increase the validity and reliability of classroom tests.

1. Develop test items that correspond with your objectives as detailed on the specification chart. Following this rule aids in matching what is tested with what was taught or, more technically, increases the content validity of the test.
2. Write the test items at a language level that all students can comprehend. If students can not read the question to know what is expected from them, then the test item is measuring something other than just what was taught.
3. Construct items so that they are independent of one another.
4. Write the test items well in advance of the test date in order to allow time for personal review and examination by a colleague. Following this practice enables the teacher to catch ambiguous or poorly stated items which otherwise might lessen the reliability and validity of the test.
5. Allow sufficient time for most student to finish all items. Lack of time pressures the student to respond rapidly which, in effect, turn out to measure something other than the specified content and behaviors.

Free-Response Items

1. Identify the higher-order processes you want the students to exhibit and be sure the essay question requires that behavior.

> What are the major differences between Newtonian and Einsteinian Physics?

Note that the student behavior might simply involve parroting previously presented information. It is conceivable that students with a good memory could respond correctly to this question without being able to apply the principles involved. Reworded, the same question can be used to elicit the higher-order behaviors desired.

> Contrast the explanations for differences in mass from Newtonian and Einsteinian Physics when '*v*' approaches '*c*'.

2. Use clear and precise language in presenting the task so that it is unambiguously defined for all students.

> Illustrate the dominant pernicious effects germinating from prodigal inhalation of noxious vapors associated with smoking.

Obviously few students would be able to provide an adequate answer to this item as stated. The problem is hidden beneath a cloak of high sounding verbiage. In general, the simplest wording possible should be used that will

convey the meaning of the question to all students. The above question might be reworded in simple vocabulary to read:

Explain the major harmful effects chain smoking has on the lungs.

3. Provide the students with a complete set of directions.

4. Require that all students answer the same set of questions. It is impossible for you to make comparative judgments of student achievement unless you have a common basis for evaluation. When students are free to select items, the commonality among responses is lost.

Restricted-Response Items

1. Be certain that a specific and well-defined problem is presented to the student by the statement or question. A student who knows the material should not be expected to guess how you want him to answer the question.

What does the coho salmon contains?

What answer would be acceptable in this instance? You might think of other answers besides 'traces of mercury' which was the intended response. Other correct responses that students might consider would include such substances such as protein, vitamins, meat, water, bones, salt, or eggs. The statement or the question must be more clearly defined for the student.

What poisonous chemical is contained in the coho salmon in trace amounts?

Note that now the students' attention and efforts are focused on a 'poisonous chemical', which eliminates the many extraneous but correct responses given for the faulty items.

2. Do not lift statements directly from the text. Since one of the primary limitations of restricted-response items involves too much focusing on knowledge level behaviors, you accentuate this weakness by encouraging rote memorization of facts when your items are direct quotes from the text.

3. In completion items omit only significant words from the statement. Omission of articles, connecting phrases, conjunctions and similar elements from a statement test the students' knowledge of insignificant details.

4. Phrase the completion item so that the blanks occur near the end of the statement. This practice enables your students to know what the question is before the blank is encountered.

5. In completion items, the blanks should all be the same length and usually scored one point each. Blanks of varying length serve as cues to the test-wise students.

6. If the question or statement requires a numerical answer, indicate to the student what units of measure are appropriate or what degree of accuracy is expected.

Structured-Response Items

True & False Questions (Alternate-Response)

1. *Use only statements which are either entirely true or entirely false.* Statements which are not absolutely true or false cause the more knowledgeable student undue difficulty in that he may be aware of possible exceptions.

2. *Avoid trivial content or trivial details that only serve to trick students.* Unless you are welling to admit that your objectives are inconsequential in nature.

3. *Beware of using specific determiners that serve as clues.*

4. *Include only one idea in each true-false item.* Complicated statements involving several ideas are often difficult to read and understand especially for younger students.

5. *Avoid use of negatives and especially double negatives.* Negatively phrased items take longer to answer and lead to more errors because the often involve a reversal of the reasoning process. These items containing negatives also test reading skills rather than competencies in the content area.

6. *Avoid ambiguous and indefinite terms that are open to interpretation*. Words and phrases such as " several," "seldom," " frequently," or "to a great extent" might be interpreted differently by each individual.

7. *Avoid length and patterning cues.* Frequently in order to make a statement unequivocally true, qualifying phrases are added that tend to increase the length of true statement. Strive to eliminate or mask the length cue by including longer items that are false. A second cue students are quick to pick up results from a set pattern of answers as TTFFTTFFTTFFT or TFTFTFT. Avoiding patterning cues is most easily accomplished if there are approximately the same numbers true and false items arranged in a random sequence.

Matching Question

1. *Use only homogenous premises and homogenous responses in a single matching exercise.* To be effective, incorrect responses must seem equally plausible to the unprepared student

2. *Use relatively short list of responses.* Ordinarily the number of responses should be more than five but less than ten or twelve.

3. *Arrange premises and responses in a logical or chronological order if possible*. Alphabetical ordering of names and numerical ordering of dates and numbers save the student time.

4. *Provide more responses than premises*. If the number of responses is equal to the number of premises, a student is able to get the last match by the process of elimination.

5. *Give the students a clear set of instructions that indicate the basis for matching.*

Multiple Choice Questions

1. *The stem should clearly formulate a specific problem*. The student should not read each of the responses before being able to decide what the question is.

2. *Keep the students' reading effort to a minimum*. This suggestion covers both parts of the MCQ, the stem and the options. Including irrelevant material might test the students' reading ability more than the behavior intended; thereby reducing the content validity of the test.

3. *Be sure only one response is considered best or correct by experts in the field.*

4. *Avoid negatively stated stems whenever possible.*

5. *Include only plausible and attractive foils as incorrect responses.* When one or more of the distractors are so obviously wrong that no body selects them, the students' chances for suggesting the right option are increased.

6. *Avoid giving students clues to the correct option.*

7. *Use the option "none of the above" or "all of the above" very rarely if at all*. Often teachers include these as a final distractor to make the number of foils in each question the same or because they can't think of any other distractor.

Improving Test Items

Analyzing Items

If you are concerned with improving your tests and instructions so that students learn more, you must resist the temptation just to file the scored tests away. Why? Well, in part because an analysis of items on the test will help you design better tests in the future.

In addition to improving the test it self, analysis of test items can also provide you with information on how well you taught particular objectives as well as aspects you need to improve. Furthermore, a postmortem examination of the test questions will guide your effort in providing corrective feedback when discussing the test with your class. This leads to improved learning and instruction. Both of these functions are based on three questions that can be asked regarding student performance on each item:

1. How difficult was the item?
2. How well did the item discriminate between student's levels of achievement?
3. In the case of MCQ, were the distractors effective in attracting students who didn't master the material?

How can you as a teacher find out all that information just by looking at a test scores? There are computer programs as well as technical methods designed specifically for analyzing tests. One of these methods is the short-cut method.

The short-cut method

A hundred test papers would take weeks to analyze. However, satisfactory results are obtained when only a portion of the total group is used. This portion consists of two groups of students: those who scored high (the high group) and those who did poorly (the low group). Each subgroup of the total class should ideally contain about 27% of the total number of students in the class who took the test.

For purposes of illustration of one method of item analysis, suppose you gave a test to a class of 35 students. The list of scores below indicates how well students scored on the test.

41	45	42	32	34	17	35
23	14	37	17	22	33	16
33	19	20	41	15	15	40
43	27	21	38	29	15	18
30	44	35	27	41	26	36

The first step in organizing the test scores for item analysis involves rearranging the scores from highest to lowest.

45	41	36	33	27	20	16
44	41	35	32	26	19	15
43	40	35	30	23	18	15
42	38	34	29	22	17	15
41	37	33	27	21	17	14

The second step consist of selecting the top 27% of the papers for the high group (H) and the bottom 27% of the papers to make up the low group (L), i.e. select 10 papers from the top and 10 papers from the bottom.

45	41	20	16
44	41	19	15
43	40	18	15
42	38	17	15
41	37	17	14
(H)		(L)	

These two subgroups of the total class will be used to compute the difficulty and discrimination indices for each item. But in order to determine these components, you need a summary of how the students in the high group and those in the low group responded to each item. The tally sheet is a useful form for summarizing this data. One advantage obtained from completing the form lies in the fact that all the information needed for obtaining the item characteristics is contained on one page. Moreover, you only have to handle the test papers for each group (H and L) once.
The procedure for obtaining the needed information involves tallying the option selected by each student for each item.

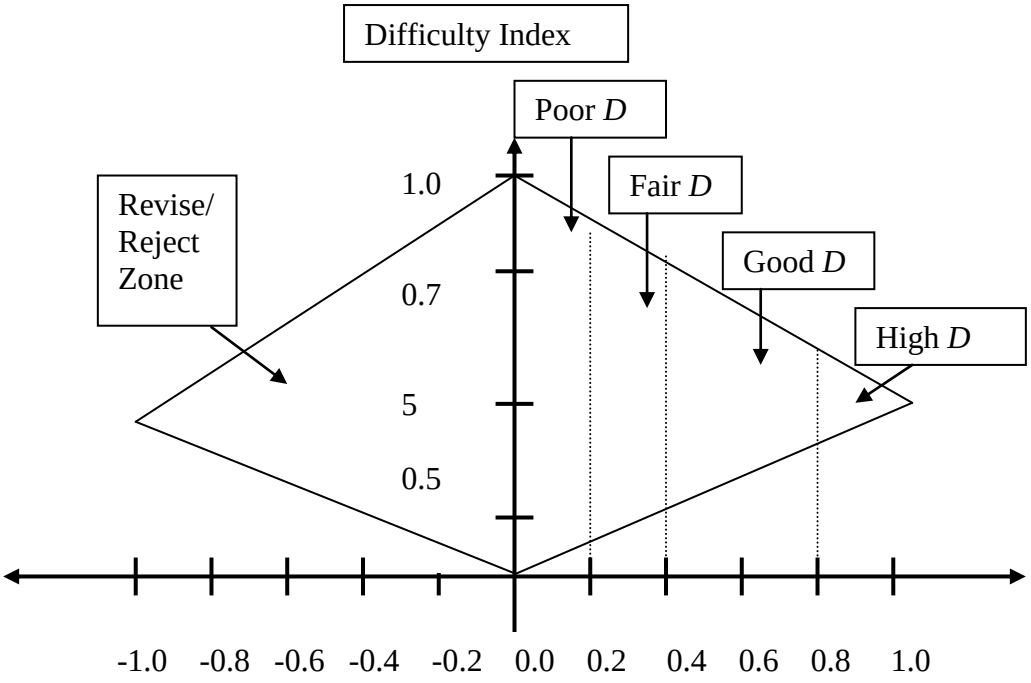
Difficulty Index – P
The difficulty (or easiness) of a test item is determined by the percent (*P*) of students who selected the right response. If over 75% of the students answered correctly, the item is judged to be an easy one. Conversely, if only 25% got the right answer, the item is considered to be a difficult one. Using the data of the high and low groups, the difficulty of the test item can be computed by adding the number right in the high group to the number right in the low group and dividing by the total number in both groups.
Note that the maximum and minimum values of *P* are 1.0 and 0.0 respectively. These values occur either when the item is so easy that everyone gets it right ($P = 1.0$) or when the question is difficult that no one responds correctly ($P = 0.0$).

Tally sheet

KEY	ITEM NO.	GROUP	OPTIONS					*P*	*D*
			A	B	C	D	E		
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							
		H							
		L							

Discrimination Index – D

Items of the same difficulty do not always discriminate equally well between students differing in achievement levels in the subject matter being tested. Hence a second test item statistic is needed in order to determine how well each item on the test differentiate between those students who do well on the test as a whole and those who score low on the test. The discriminatory power of a test item can be described as the difference between the percent of students in the high group answering an item correctly and the percent of students responding correctly in the low group. The discrimination index may be computed by dividing the number of right responses in the high group by the number of students in the high group then subtracting the number of right responses in the low group divided by the number of students in the low group.


Difficulty Index
Poor D
Fair D
Good D
High D
Revise/
Reject
Zone
1.0
0.7
5
0.5
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
Discrimination Index