

INFORMAL TESTS FOR CLASSROOM USE

W.B. Elley,
Assistant Director
New Zealand Council for Educational Research.

WHY TEST?

In the course of his career, every teacher will have occasion to prepare, administer and mark hundreds of tests of his pupils' attainment. Sometimes these tests will consist of carefully selected formal written exercises with a rigid time limit and an elaborate marking system; sometimes they will be spontaneously constructed, orally presented and evaluated only superficially. But all will have these characteristics which distinguish them from external examinations and standardized tests:

- (i) Classroom tests are prepared by the teacher or headmaster for local rather than national use. They are usually prepared at short notice, without the benefit of special expertise, panel review, or pretesting of questions.
- (ii) Classroom tests are usually designed to evaluate the pupils' mastery of short units of work recently studied, or objectives which are specific to a school, a class, or a lesson. A teacher may prepare a short test on multiplication and division with decimals, or the causes of World War II, or a list of irregular French verbs. By contrast, an external examination usually evaluates the fruits of a year's study - or more, while a standardized test normally samples basic objectives which are developed over an extended period of time, and are not dependent on the teaching of a particular course.

However sophisticated his test preparation procedures, it should be apparent to all that a teacher requires the results of classroom tests to make decisions about his teaching:- whether to proceed or to back track, whether to change a teaching method, or to introduce a new topic. Judgements have to be made about classification and selection of pupils, advice must be offered about course changes and vocational plans, suitable materials and approaches must be found for children at all levels of ability, pupils must be identified for special treatment. Whether a teacher's decisions are required for groups or for particular individuals, they are more likely to be sound if they are based on accurate information about the abilities and attainments of his pupils. If this information is to be helpful, it should be obtained from tests which are both reliable and valid. Tests which are too easy or difficult, tests which are too short or too long, tests which sample only part of the course, or which weight certain parts too heavily, tests which are ambiguous in their directions, or which leave too much to chance, tests which cannot be marked with reasonable objectivity - such tests may mislead both teacher and pupil, confirm erroneously-held prejudices, and occasionally lead to injustices with far-reaching effects. How can teachers prepare classroom tests which will produce results in which they can place confidence? What test construction methods are likely to produce tests of adequate reliability and validity? First we must examine these criteria of a good test. What does it mean to say that a test is reliable and valid?

RELIABILITY.

Tests are reliable if they produce consistent results, if they produce similar marks on different occasions. If a pupil gains 100% in a foreign language dictation test today, and only 50% tomorrow, then the results are not consistent, the tests are not sufficiently reliable to base judgements on. If a pupil is placed first in his class in a test of multiplication and division of decimals on one occasion and is 20th in a subsequent test of the same skills, we can conclude that the tests are not reliable indicators of his ability.

To be reliable a test must normally be long enough to minimize the effects of chance factors in the content and skills included in the test. With a short test, a pupil may be lucky, because he happened to know or guess correctly the few questions that were asked, whereas he knew very little about the areas untouched by the test. A standardized test of reading, mathematics or language, normally requires at least 40 sound objective-marked questions to reach a satisfactory level of reliability. To make decisions about individual pupils, a teacher-made test will probably require more questions than this. For judgements about groups, a teacher may get by with fewer. Just how long a particular test should be depends on the type of material tested, the amount of supplementary information available, and the importance of the decisions being made. Thus a test of a highly specific skill, such as arithmetical computation, or typing, may produce reliable results within ten minutes. If however, we wish to examine a pupil's grasp of mathematical relationships, or his understanding of a period of history and to make decisions about future schooling on the basis of the results, we may wish to extend the test over two hours to gain maximum reliability. For such skills as essay-writing ability, or oral expression, it is commonly found that pupils vary so much in their performance from day to day that the only way to gain adequate reliability is to test the pupils on several topics (over several occasions), and to combine the marks given by two or three independent markers.

Other requirements of a reliable test are clear, precise directions and reasonable time limits. The questions should be unambiguous, neither too easy nor too difficult; they should discriminate well between good and poor pupils, and they should be capable of reasonably objective scoring.

VALIDITY

A good test must be valid. This means that, in addition to measuring a pupil's attainments reliably, it should be relevant to the needs of the tester. It should cover the unit or course adequately, sampling each content area and skill in appropriate proportions. If a teacher knows precisely what his objectives are, he can usually tell, by analysing the questions of a test, whether they conform closely to the objectives he has adopted i.e. whether the test is valid for his purposes.

To illustrate, a 100-item test of mathematical computation may be highly reliable, and yet be quite invalid for measuring achievement in a course of modern mathematics which emphasizes concepts, relationships and reasoning. The objectives of the test do not match the teaching objectives. Again, a test of geography which focusses on isolated details about populations, areas, climate, exports, capital cities and the like, would produce irrelevant results for a teacher who stressed broad concepts, generalized skills and underlying relationships. A valid test of such

objectives may require novel or fictitious situations on which to base questions so that a pupil can demonstrate that he has attained these objectives, regardless of the particular factual details he has acquired.

To ensure maximum validity for his tests, then, it is important for a teacher to spell out, as clearly as possible, precisely what his objectives are, and to build his questions around these, in the appropriate proportions. Tests which develop without such planning often degenerate into factual quizzes of the low-level, isolated, easily testable fragments of the course.

DEVELOPING THE TEST.

- (i) Once a teacher has decided on the purpose for his test, he should consider the various objectives he has in mind, and how he might best classify them. For a content-oriented course, such as science or history, he might first divide the course into the main content areas, and ensure that each receives a fair ratio of questions. A general science course may be classified into three main areas - say chemistry, physics and biology. A more specific classification, for a biology course, might be living organisms, life processes, conservation, heredity and reproduction, and evolution. In addition a teacher should ensure that questions test different levels of understanding. Some tests concentrate on examining for recall of specific information, some for understanding of important ideas, some for application to new situations, and so on. In a science test, a useful classification system for the objectives might be
 - (a) Knowledge of facts and conventions
 - (b) Understanding of concepts and principles
 - (c) Ability to apply the scientific method to new problems
 - (d) Knowledge of industrial applications.In language subjects the content areas are less easily defined and it may be more appropriate to classify the objectives of the course according to the skills to be tested - reading, writing, translation, dictation, etc.
- (ii) Once objectives are classified, a blueprint or table of specifications can be drawn up which sets out the content areas and the objectives, and allows them to be weighted on some rational basis, before the test questions are prepared. For maximum validity, a test will normally weight most heavily those topics or objectives which have been given most emphasis in the course or unit taught. But all areas should be tested where possible.

An example of a fictitious test blueprint is set out below.

Sample Blueprint for a Mathematics Examination

Objectives	Nos.& Numerals	Measurement	Fractions	Geometry	Total
Knowledge of terms, facts	10	5	5	5	25
Understanding of concepts	10	5	5	10	30
Routine calculations ..	5	5	5	0	15
Application to new problems ..	15	5	5	5	30
Total emphasis ..	40	20	20	20	100

(iii) The third stage in developing the test requires a decision on the form of the questions to be asked. There is no question type ideally suited for all purposes. For instance, short-answer questions which require pupils to fill in the blank or complete a sentence are useful for covering a wide range of facts in a short time. Outside of mathematical subjects they are less useful for estimating depth of understanding without introducing some ambiguity in the question or subjectivity in the marking. Multiple-choice questions are widely used in standardized tests and external examinations because they can sample the whole course widely and efficiently, and test higher objectives, but such questions are not easy for classroom teachers to prepare and they do not examine the ability of the pupil to generate and organize his own ideas. Matching questions are best suited to measuring knowledge of homogeneous sets of facts or conventions. Pupils may be asked to match books with their authors, chemical compounds with formulae, countries with exports, etc. They should not be used however, unless the contents of each list form a homogeneous group, so that each item on one list is a plausible match for each item in the other list. True-false questions may have some value in classroom tests since they enable the teacher to sample widely in a short time, but they are frequently superficial, they are unsuitable if the truth of each statement is not absolute, and they are prone to be unreliable due to guessing on the part of the pupils. They can be modified of course, by requiring pupils to correct false statements, or to classify a statement as "sometimes true", depending on other factors. Perhaps their greatest value is as a starting point for classroom discussion. Essay questions compensate for some of the deficiencies of other question types in that they do require the pupil to express his own ideas, and to demonstrate fluency and organization. However, they cannot measure as many aspects of a course as do short-answer questions, and they are

difficult to mark reliably.

Before deciding on the kinds of questions to use then, a teacher should consider the various pros and cons outlined above, in relation to his own expertise in item-writing, the number of pupils involved, the time available for setting and marking, the degree of reliability required and the kinds of decisions to be made with the results.

(iv) Preparing the questions to fit the blueprint is the fourth stage. Here there are many pitfalls, and no short-cuts to success. So often when questions are hurriedly prepared they turn out to be ambiguous, too easy, too difficult, or unsuitable for some other reason. The following checklist may alert teachers to the kinds of weaknesses likely to be found in their questions.

(a) - General

- Keep questions brief, simple and free from complex verbal instructions, double negatives etc.
- Test only important facts and skills; avoid trivia, catch questions, and irrelevant material.

(b) Completion Questions

- Use a single blank in each question.
- Place the blanks near the end of the sentence.
- Ensure that there are a finite number of correct answers.
- Make all blanks approximately the same length.

(c) Multiple-Choice Questions:

- Use only plausible distractors.
- Ensure that there is only one acceptable answer.
- Avoid the stereotyped language of textbooks in the correct answer.
- Beware of grammatical clues and verbal associations which help the uninformed.
- Make the correct option the same length as the distracting option.
- Avoid overlap in the options.
- Avoid any discernible pattern in the correct answers.

(d) Matching Questions:

- Clarify the instructions so that pupils know the basis for matching.
- Use only homogeneous sets of items in each list.
- Make an unequal number of items in each list.
- Use fewer than ten items in each list.

(e) Essay Questions:

- Ensure that the question as it is asked cannot be more adequately measured by another approach.
- Structure the question in such a way that pupils know what to include, what to omit, and how much to write.
- Ask several short questions of different types rather than one long question.
- Avoid giving pupils a choice of questions unless it is absolutely necessary.
- Prepare a model answer before the test, but be prepared to revise it in the light of pupils' answers.

These principles may not always be applicable or even justifiable, but they do point to frequent sources of weaknesses in classroom tests. Such weaknesses can often be overcome by working with a colleague or panel of teachers, by requiring somebody to answer the questions while the test is being prepared, or by pre-testing the questions on a sample of pupils similar in ability to those for whom the test is designed.

There are no perfect paper-and-pencil tests. All are somewhat artificial; all are subject to pupil fluctuation in concentration; all provide only a sample of pupil knowledge; all therefore are to some extent unreliable. A close observance of the principles outlined above, however, should help teachers to polish up their testing procedures. Further improvements can be effected by studying textbooks on the subject, by examining well-constructed standardized tests, by item analyzing one's own tests, and by discussing one's efforts with other teachers.