Quantitative research methods in educational planning

Series editor: Kenneth N.Ross

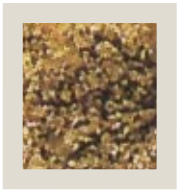
Module 6

John Izard

# Overview of test construction

UNESCO International Institute for Educational Planning

# Content



# Assessment needs at different levels of an education system

1

Assessment of student learning provides evidence so that educational decisions can be made. We may use the evidence to help us evaluate (or judge the merit of) a teaching programme or we may use the evidence to make statements about student competence or to make decisions about the next aspect of teaching for particular students.

The choice of what to evaluate, the strategies of assessment, and the modes of reporting depend upon the intentions of the curriculum, the importance of different parts of the curriculum, and the audiences needing the information that assessment provides. For example, national audiences for this information may include both those who will be making decisions and those who wish or need to know that appropriate decisions have been taken.

Educational decisions which require information about the success of learning programmes, or which require information about which students have reached particular levels of skill and knowledge, depend upon valid (and therefore reliable) measures to inform those who make the decisions. The type of information will depend upon whether the decisions are being made at the personal, school, regional, or national level. Variables which are seen to influence the outcomes of education may, or may not, be within the province of school systems to alter. For example, socio-economic circumstances are known to have influences on student achievement, but teachers are not generally able to change the

socio-economic circumstances of the families in their school's community. By contrast, other variables are able to be manipulated to produce changes in student achievement (We say these variables are *malleable*). For example, better teacher in-service training and the provision of improved instructional materials can improve student achievement.

In order to measure progress, tests need to be given more than once so that changes can be identified. For example, to assess the impact of new programmes to improve schools, baseline measures are needed to describe the effectiveness of the teaching provision before the innovation, so that subsequent measures can be used to judge the effectiveness of the implemented innovation.

## Student, teacher and parent assessment needs

At the individual student level, students, teachers, and parents need information about student performance expressed in ways which not only identify strengths and weaknesses, but which also suggest what might be done to capitalise on the strengths and to overcome the weaknesses. Assessment data can only be understood in the context in which they were collected. For example, a score of 59% is meaningless without knowing what teaching/learning situations have been provided, how long the educational programme has been offered, whether the student has actually been present for all or most of the programme, what questions were asked, and what answers were expected. Such a score also has implicit messages about precision – the accuracy is implied to be to the nearest half of a percentage point, although such precision is very rarely achieved in educational assessment.

## School level assessment needs

At the school level, the school principal and senior administration group generally require information about classes rather than individual students. This information may be expressed in association with information from classes in other schools in the district, region, or nation. Such comparisons generally concentrate mainly on relative standing. The relative standing of a particular school may improve for reasons that are not related to the skills of the teachers or the educational programme of the school. For example, relative standing may improve because schools select pupils who will do well even if the teaching is poor. Rather than concentrating on relative standing, it is better to focus on information expressed in terms of expected learning levels and progress towards educational goals. Then actions taken can relate to ensuring that accepted educational goals will be met for all students in the school. In this case success would be judged by taking into account the extent to which a school has ensured that every student has made good progress.

## Regional and national assessment needs

At the regional level (including state and provincial levels) the information required is generally concerned with improving the effectiveness of larger numbers of schools. Evidence of school achievement might be based on a wider range of indicators, such as effective use of resources provided to a school, provision of educational programmes which meet policy guidelines, and the extent to which the community where the school is placed is involved in the educational programmes.

At national level the information required must relate more to policy issues, national planning, and the resource implications for competing options in educational plans.

*It is particularly important for National officials to be sensitive to long-term trends in their education system's capacity to assist all students to make progress towards achieving a high standard of physical, social and cognitive development. In some circumstances these trends will call for intervention in what is seen as an emerging and widespread inability of students to achieve success in a specific part of the curriculum. In other circumstances, the focus will be on the curriculum itself because it may be seen as being in need of revision and restructuring in order to take account of recent research and/or new social and economic conditions.* (Somerset & Eckholm, 1990, p.18)

Those who are taking action should also know the likely direct and indirect effects of various action options, and the costs associated with those options. They will include politicians, high level advisors, senior administrators, and those responsible for curriculum, assessment, teacher training (pre-service and in-service), and other educational planners.

That is, those taking action need to be able to provide evidence that their actions do 'pay off'. For example. politicians have to be able to convince their constituents that the actions taken were wise, and senior administrators need to be able to show that programmes have been implemented as intended and to show the effectiveness of those programmes. It is important for such officials to realise that effecting change requires more than issuing new regulations. At the national level, action will probably be needed to train those responsible for implementing change.

# What is a test?

2

One valid approach to assessment is to observe everything that is taught. In most situations this is not possible, because there is so much information to be recorded. Instead, one has to select a valid sample from the achievements of interest. Since school learning programmes are expected to provide students with the capability to complete various tasks successfully, one way of assessing each student's learning is to give a number of these tasks to be done under specified conditions. Conventional pencil-and-paper test items (which may be posed as questions) are examples of these specially selected tasks. However other tasks may be necessary as well to give a comprehensive, valid and meaningful picture of the learning. For example, in the learning of science subjects practical skills are generally considered to be important so the assessment of science subjects should therefore include some practical tasks. Similarly, the student learning music may be required to give a musical performance to demonstrate what has been learned. Test items or tasks are samples of intended achievement, and a test is a collection of such assessment tasks or items.

Single, discrete items may not be reliable (or consistent) indicators of achievement. However, when a number of similar items or tasks are combined as a test, we can look at patterns of success on the test. Such patterns tend to be more dependable indicators because they are based on multiple sources of evidence (the various separate assessment tasks).

Clearly, the answer for one item should not depend on information in another item or the answer to another item. Otherwise this

notion of combining independent pieces of evidence would be lost. (The same idea extends to other tasks where results are combined with test results to document learning achievement.)

This approach of giving all students of a particular age the same sample of assessment tasks is of value at both the individual student and school levels. Teachers and school principals can examine two types of profile to evaluate delivery of educational programmes. The teacher might look at the performance of individual students in each of the areas assessed in order to find out more about the extent of progress since the previous assessment. The teacher and the principal might look at the performance of the assessment tasks themselves to identify those topics presenting special difficulties within school classes or across classes.

Regional and national officials may wish to review performance on particular assessment tasks also, but the large volume of data and the expensive resources needed to collect the data, process and interpret it, preclude collection of this information on an individual basis for every student. Well-designed probability samples of students will provide more economic and quite accurate ways of estimating regional or national performance. Recent advances in testing technology mean that students in these samples need not attempt identical test questions – and therefore the 'coverage' of the collected information can be extended to a wider range of topics. When such data are collected it is important to ensure that information is gathered on variables that are influential but which (for schools) are not malleable as well as on variables which can be influenced by schools. With care, regional and national officials can take (statistical) account of the non-malleable factors when assessing the impact of variables that can be influenced by schools.

Generally traditional examinations are *not* appropriate for reviewing regional or national performance over a period of time. National examinations cannot show the extent of *improvements* of teaching skill, the extent to which all parts of the curriculum

are working, or the *magnitude of improvement* which results from deployment of resources as a result of policy changes. The scoring of national examinations which are multiple-choice in format is likely to be consistent, but such questions are not likely to be used in a recurring pattern because of the potential for breaches of test security in such a high-stakes assessment context. The scoring of open-ended and short-answer format questions will include variation due to scorer behaviour as well as that due to candidate behaviour, and the same examiners are not likely to assess a comparable question in subsequent years.

These problems in traditional examining mean that assessing changes (in regional or national performance over time) requires a range of specially developed low-stakes tests. In the development of these tests, care needs to be taken that the questions used on one occasion are comparable to those used on another, even though they are not the same questions. This comparability must be demonstrated empirically at some stage. Usually this means that both sets of questions are given to another sample of students representative of the range of achievement. Then questions apparently similar with respect to content and coverage can be checked to see whether students respond to the questions in a comparable way. Questions that are comparable will have similar ranges of difficulty, will reflect similar performance by significant sub-groups of the population (such as males, females, ethnic minorities, city and rural), and will have similar discrimination patterns over the range of achievement. (In other words, low achievers will have similar performances on both sets of questions, middle level achievers will have similar performances on both, and high achievers will have similar performances on both. These specially developed tests can be used with relatively small representative samples to assess the extent of changes for the purposes of monitoring effects of additional funding, changes in the provision of teachers, or the effects of introducing new instructional materials.

# 3 Interpreting test data

## The matrix of student data

When data from a test are available, the requirements of the various audiences interested in the results differ. This can be illustrated using the matrix of information shown in *Figure 1*.

The students and their parents will focus on the total scores at the foot of the columns. High scores will be taken as evidence of high achievement, and low scores will be taken as evidence of low achievement. However, in summarising achievement, these scores have lost their meaning in terms of particular strengths and weaknesses. They give no information about which aspects of the curriculum students knew and which they did not understand. Teachers, subject specialists, curriculum planners, and national policy advisors need to focus on the total scores shown to the right of the matrix. These scores show how well the various content areas have been covered. Low scores show substantial gaps in knowledge where the intentions of the curriculum have not been met. High scores show where curriculum intentions have been met (at least for those questions that appeared on the test).

**Figure 1. Matrix of student data on a twenty-item test**

| | Students | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Items** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | |
| **1** | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | **15** |
| **2** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | **14** |
| **3** | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | **14** |
| **4** | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | **11** |
| **5** | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | **13** |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | **11** |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | **8** |
| **8** | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | **13** |
| **9** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | **9** |
| **10** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | **7** |
| **11** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | **7** |
| **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | **6** |
| **13** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | **7** |
| **14** | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | **11** |
| **15** | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | **12** |
| **16** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | **8** |
| **17** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | **8** |
| **18** | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | **6** |
| **19** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | **4** |
| **20** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | **5** |
| **21** | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | **10** |
| | **3** | **4** | **5** | **7** | **7** | **9** | **10** | **10** | **12** | **12** | **14** | **14** | **14** | **14** | **15** | **15** | **17** | **17** | **199** |

## Designing assessments to suit different purposes

Traditional examinations give every student the same task so that individuals can be compared, for example, by comparing the column totals in *Figure 1*. As time is limited and the cost of testing large numbers of candidates is high, the number of tasks used has to be relatively small. As the costs of assessment are roughly proportional to the number of cells in the matrix, the number of questions asked in traditional examinations will be limited to contain costs. The resulting matrix will be wide to cater for many students but not very deep because of the limited number of test items (see *Figure 2*).

**Figure 2. Traditional examination data matrix**

| | Students | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Items** | **1** | **2** | **3** | **4** | **5** | • | • | • | • | • | • | • | • | • | |
| **1** | | | | | | | | | | | | | | | |
| **2** | | | | | | | | | | | | | | | |
| • | | | | | | | | | | | | | | | |
| • | | | | | | | | | | | | | | | |
| • | | | | | | | | | | | | | | | |

Information about many important issues cannot be collected because so many students have to be tested. Traditional examination questions are a sample of those assessment tasks that can be given to all students in a convenient format and they usually ignore those assessment tasks that cannot readily be given to all students. By contrast, tests used in national assessments are likely to differ from the usual public examinations. National assessments, which gather information on a much larger number of topics, will

need to limit the number of students to contain costs. The resulting matrix will be narrow, but very deep because of the larger number of test items (see *Figure 3*).

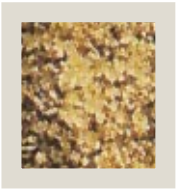
National planners do not need to know every detail of every individual's school performance. Just as a medical practitioner can take a sample of tissue or body fluids under standard conditions, subject the sample to analyses, and draw inferences about a person's health, a national assessment can take a sample of performance by students under standard conditions, analyse the data and draw inferences about the health of the educational system.

If the sample of evidence is not appropriate or representative the inferences made about current status in learning will be suspect, regardless of how accurately the assessments are made.

**Figure 3. National assessment test matrix**

| | Students | | | | |
|---|---|---|---|---|---|
| **Items** | **1** | **2** | **3** | **4** | **5** |
| **1** | | | | | |
| **2** | | | | | |
| **3** | | | | | |
| **4** | | | | | |
| **5** | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |
| • | | | | | |

If the number of questions in a national assessment is large, the testing required may be more than can be expected of typical students. It will be necessary to choose more than one representative sample of students so that evidence can be gathered on each important issue.



# 4 Inferring range of achievement from samples of tasks

Assessment involves selecting evidence from which valid inferences can be made about current status in a learning sequence. If we do not select an appropriate sample of evidence the conclusions we draw will be suspect, regardless of how accurately we make the assessments. It is possible to make consistent assessments which are not meaningful in the context of the decisions we must make.

For example, we could weigh the students and give the highest scores to those with the largest mass. This assessment could be very consistent, particularly if our scales were accurate. However, this assessment information is not meaningful when trying to judge whether learning has occurred. To be meaningful in this context, our assessment tasks (items) have to relate to what students learn. The choice of what to assess, the strategies of assessment, and the modes of reporting depend upon the intentions of the curriculum, the importance of different parts of the curriculum, and the audiences needing the information that assessment provides.

## Choosing samples of tasks

Tasks chosen have to be representative so that:

- dependable inferences can be made about both the tasks chosen for assessment and the tasks not chosen;

- all important parts of the curriculum are addressed;
- achievement over the full range is assessed (not just the narrow band where a particular selection decision might be required on a single occasion).

Choosing representative tasks may be difficult. Remember that the tasks have to represent the whole curriculum, not just the parts that can be tested with pencil-and-paper test items. Further, some pencil-and-paper test items are better than others in assessing interpretation and understanding, and in providing information in a form that can be used to make teaching decisions. If these qualities are required by the curriculum then they have to be assessed by representative tasks. Pencil-and-paper test items which only require memory are much easier to write but cannot provide other essential evidence. For example, being able to give the correct answers to number facts such as 6+3=?, 9+5=?, and 7x3=? does not provide sound and dependable direct evidence about whether a student can read a graph, or measure the length of a strip of wood. If reading graphs is important, then the assessment tasks should include some that involve the reading of graphs. If measuring lengths is important, then length-measuring tasks must be used to decide whether this curriculum objective has been met. Constructors of tests often draw up a list of topics and types of skill to specify what the test should cover.

On subsequent occasions it may be necessary to choose different representative tasks (otherwise the first group of tasks tested may be the only ones taught and therefore will no longer be representative of all the curriculum). The tests have to include easier tasks as well as more difficult tasks. The easier tasks will allow students to show more of what they have learned. The more difficult tasks will allow the best students to show where they excel.

The function of the tasks is to provide meaningful evidence. The tasks have to be matched in difficulty (and complexity) to the level of the students who are going to attempt the tasks. Tasks may be too difficult. If students do not engage with the tasks, little or no evidence is provided.

If students cannot do any of the tasks then they cannot provide any evidence of their achievements. If two such assessments are made it will appear that these students have not learned anything (because there will be no change of scores) even though they may have learned a great deal of important knowledge and skills. Such assessments are faulty in that they fail to recognise learning that has occurred. (A test with items which do not allow the less able students to show evidence of their learning may be referred to by saying that the test has a 'floor' effect.)

Tasks may also be too easy. If all students can do all of the tasks then the most able students will not be able to provide evidence of their advanced achievements. If two such assessments are made it will appear that these able students have not learned anything (because their scores cannot improve) even though they may have learned a great deal of important knowledge and skills. Such assessments are also faulty in that they fail to recognise learning that has occurred. (A test with many easy items which do not allow more able students to show evidence of their learning may be referred to by saying that the test has a 'ceiling' effect.)

The range of complexity of tasks should be at least as wide as the expected range of achievement for the students being assessed if evidence of learning is required about all students. Writing test tasks and items with desirable properties requires a great deal of skill over and above knowledge about the curriculum, and about how students learn. A team of trained item writers can usually produce a better range of items to consider for trial than any individual (or group of individuals working alone). Item writing

without the benefit of interaction with colleagues is generally inefficient and tends to be too idiosyncratic, representing only one person's limited view of the topic to be assessed. When inspiration is lacking, the items written may degenerate to a trivial level.

## The wider implications of choosing samples of tasks

Assessment has considerable influence on instruction. Topics chosen for assessment and the items chosen for those topics convey to students a view of what is considered important by those who make the assessments. Where the assessments are external to the schools, the items chosen convey a similar message to teachers as well. Conversely, topics or items not chosen indicate what is considered not important.

When assessment results have high stakes (as in the case where results are used to select a small proportion for the next stage of schooling or for employment), the chosen assessment tasks have a high degree of influence on curriculum, teaching practice, and student behaviour. When public examination papers are published, teachers and students expend a great deal of effort in analysing these papers, to practice test-taking skills, and to attempt to predict what topics will be examined so that the whole curriculum does not have to be studied. These practices of restricting learning to examinable topics may lead to high scores being obtained without the associated (and expected) coverage of the intended curriculum.

Narrow testing practices have undesirable influences on teaching. For example, tests which encourage memorizing facts rather than understanding relationships lead to teaching which ignores understanding regardless of national needs for people with such

skills. One possible consequence is for education authorities (and the general public) to lose confidence in the examination system because individuals with high scores do not have the skills and understanding required. Note that when this happens, the authorities often increase the score for a pass hoping that this will help regain public confidence. However requiring a higher score on an inadequate test cannot solve a problem which depends on more relevant items being asked. The only solution is to improve the quality of the assessments so that they match the curriculum intentions.

This may cost more at first because the levels of skill in writing such tests are much higher. Generally teams of item-writers are required rather than depending upon a very limited number of individuals to write all of the questions. The pool of experienced teachers with such skills will not increase in size if teachers are not encouraged by the assessment procedures to prepare students for higher quality assessments. Further, item writing skills develop in part from extensive experience in writing items (especially those that are improvements on previous items). Such experience of item writing, of exploring the ways that students think, and of considering the ways in which students interpret a variety of evidence, is gained gradually. Many good item writers are experienced and sensitive classroom teachers who have developed the capacity to construct items which reveal (correct and incorrect) thought processes.

Examination assessment may fail to sample tasks where there are multiple solutions, where problems may be solved in different ways, or where more than pencil-and-paper skills are required. For example, pencil-and-paper examinations are not good measures of practical tasks, of a person's skill in working in a group, or of the capacity to develop alternatives if the first attempt fails to work. Yet in many societies (if not all) being able to do practical tasks, to work in groups, and to solve local problems are skills considered to be

essential for survival and community participation. If an excessive emphasis is placed on pencil-and-paper assessments then the result may be to devalue some kinds of valuable skills that are essential in many communities.

# 5 What purposes will the test serve?

Test results are interpreted in many ways. One important way involves comparing each student's score with the scores of a group of students (who are supposed to be like the student for whom the comparison is being made). Such comparisons can tell us how well one student scored relative to another but they do not tell us which students are competent in a chosen area or suggest what might be done to increase performance.

A second important way of interpreting results involves comparing each student's results with a set of fixed requirements. Such comparisons can tell teachers and administrators the proportion of students with acceptable levels of skill and can identify those topics needing extra or different teaching and/or learning strategies. Tests prepared for this second purpose can be used to rank-order students as well. (Tests prepared for rank-ordering tend to exclude questions which many can answer successfully because there is little interest in what skills people have or do not have if the only purpose is to establish an ordered list of students.)

Describing changes in terms of total scores only is counterproductive. The scores can only be understood in the context in which they were collected. For example, a score of 60 per cent on one test (with easy items) may be worth less in achievement level terms than a score of 30 per cent on another test with difficult items in the same content area.

## Results used to compare students

**Norm-referenced tests:** These tests provide the results for a reference group on a representative test and therefore scores on the test are normally presented in terms of comparisons with this reference group. If the reference group serves as a baseline group, norm-referenced scores can provide evidence of learning improvement (or decline) for the student population although this is in terms of a change in score rather than an indication of what students can now do (or not do) compared with what they could do before. If changes in score are reported (for example, a difference in average score), administrators have little evidence about the strengths and weakness reflected in the results for particular topics and may rely on rather limited experience (such as their own experiences as a student) to interpret the changes. This could result in increased expenditure on the wrong topic, wasting scarce resources, and not addressing the real problems.

This evidence may be compromised by the actual test becoming known to teachers. The result may be that they (quite naturally) begin to emphasize the work covered in the test and therefore scores may well rise. This rise does not provide evidence of improved performance on the curriculum as a whole by teachers and students. Where the rise is at the expense of studies in other important parts of the curriculum not sampled in this particular test the effect is to destroy the representative nature of the actual test as a measure of progress in the curriculum.

The use of norm-referenced tests also depends on the curriculum remaining static. If curriculum changes are introduced or time allocations are changed, a representative 'snapshot' of the initial curriculum may not be representative of the changed curriculum. Comparisons with the original reference group are then not appropriate.

Norm-referenced tests are often wide-range tests. They can be used to provide an order of merit for competitive selection purposes where the test chosen is relevant, in general terms, for the skills needed. However the scores which are used to determine each candidate's standing provide little information of direct use to a teacher. Other information is required with the test if advice to students and teachers is considered to be one of the important outcomes associated with using the test scores. Such tests are often prepared with a particular curriculum in mind. It is important to check each question against the curriculum to see whether there is a good match between the curriculum and the norm-referenced test authors' 'assumed curriculum'. For example, Australian mathematics syllabuses introduce algebra and geometry in the seventh year of schooling while some curriculum statements from Northern America assume a much later introduction. Further, the balance of items for the curriculum may not match the balance shown in the chosen test. If some items in the chosen test are not appropriate for the curriculum then the comparison tables of total scores for the chosen norm-referenced will not be appropriate or meaningful. Some test publishers are able to re-calculate norm-referenced tables for meaningful comparison purposes, but this often depends on the availability of the full data and trained staff who are able to undertake the re-calculation.

Very few tests provide the user with a strategy for making such adjustments for themselves although some tests prepared using Item Response Theory or Latent Trait Theory do enable qualified and experienced users to estimate new norm tables for particular sub-sets of items.

## Results used to compare students with a fixed requirement

**Criterion-referenced tests:** These tests report performance in terms of the skills and knowledge achieved by the students and do not depend explicitly on comparisons with other groups of students. Often a criterion referenced test will include all of the criteria in the curriculum that are of importance (rather than rely on a sample as in the case of norm-referenced tests).

Criterion-referenced scores can provide evidence of learning improvement (or decline) of the student population as an indication of what students can now do (or not do) compared with what they could do before. This evidence may be reported as proportions of students who have achieved particular skills and is less susceptible to curriculum changes (provided those skills are still required in the changed curriculum). There is less likelihood of criterion-referenced tests being compromised by the actual test becoming known to teachers. If they emphasize the work assessed by each test (rather than particular items being used for a test) they will have covered all important objectives of the curriculum. A rise in the proportion of successful students will provide evidence of improved performance on the curriculum as a whole by teachers and students (provided that the rise was not achieved by excluding students on the basis of school performance or by being more selective in enrolling students).

**Mastery tests:** These tests are generally criterion-referenced tests with a relatively high score requirement. Students who meet this high score are said to have mastered the topic. It is assumed that the mastery test has sufficient items of high quality to ensure that the score decision is well founded with respect to the domain of interest.

For example, in mathematics the domain might be 'all additions of pairs of one-digit numbers where the total does not exceed 9'. A mastery test of this domain should have a reasonable sample of all possible combinations of one-digit numbers because the mastery decision implies that all can be added successfully even though all are not tested. This simple example should not be taken to imply that mastery testing is limited to relatively trivial skills. A more complex example is the regular testing of airline pilots. Safety requirements result in high standards being set for mastery in many areas. Failure to reach mastery will result either in further tuition under the guidance of an experienced tutor or in withdrawal of the permission to fly.

## Other labels for categories of test use

**Speed test and power test:** Some tests have very easy items but there is a limited amount of time to answer them. Such speed tests are used to see how quickly students can work on skills they have already mastered. One example is a test of keyboard skills. The teacher may wish to find out how fast students can maintain accurate work when typing data on a typewriter or computer keyboard.

In contrast, power tests are concerned with identifying skills which have been mastered. Power tests require adequate samples of student behaviour – so having sufficient time to attempt most of the items is an essential pre-requisite for such tests.

**Aptitude or ability test and achievement test:** Achievement tests may be used to assess the extent to which curriculum objectives have been met in an educational programme. Such tests should have tasks which relate to the learning that students have to demonstrate. Since future learning depends to some extent on past learning, success on such achievement tests may provide evidence

of future success (provided that other conditions such as good teaching, adequate health care, and stable family circumstances are maintained).

Tests which are constructed specifically to gather evidence about ability to learn are referred to as aptitude or ability tests. Results on such tests are used to predict future success on the basis of success on the specially selected tasks in the aptitude test. Often these tasks differ from the usual school learning requirements and depend to some extent on learning beyond the school curriculum. Of course, teaching students the test items and the corresponding answers may result in an increase in score without actually changing a student's (real) aptitude.

**Objective test:** The term 'objective' can have several meanings when describing a test. It can mean that the score key for the test needs a minimum of interpretation in order to score an item correct or incorrect. In this sense, an objective test is one which requires task responses which can be scored accurately and fairly from the score key without having knowledge of the content of the test. For example, a multiple-choice test can be scored by a machine or by a clerical worker without either the machine or the clerical worker having had to reach a high level of expertise on the material being tested.

A less common usage relates to the extent of agreement between experts about the correct answer. If there is less argument about the correct answer the item is regarded as more objective. However the choice of which items (whether objective in their answer format or not) are to appear on a test is subjective, in that it depends on the personal preferences and experiences of those constructing the test.

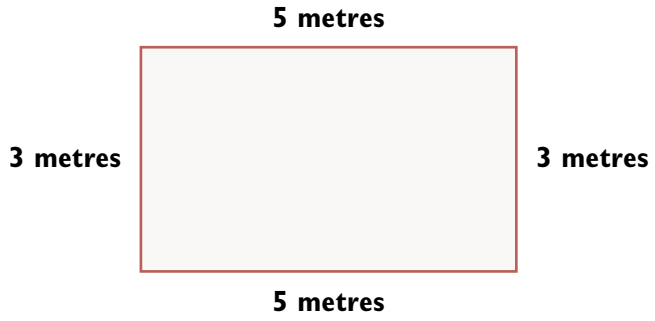
**Standardised test:** The term 'standardised' also has a number of meanings with respect to testing. It can mean that the test has an agreed format for administration and scoring so that the task

is as identical as possible for all candidates and there is little room for deviation in the scoring of candidate responses to the tasks. Another meaning refers to the way in which the scores on a tests are presented. For example, if scores are given as a raw score divided by some measure of dispersion like the standard deviation, the resulting score scale is said to be in terms of standardised scores (sometimes called standard scores).

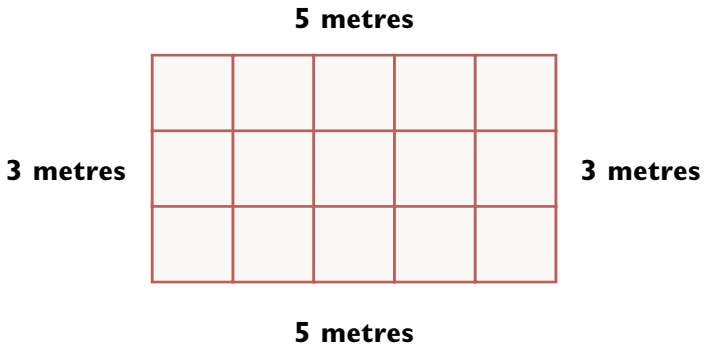
Finally, the term can refer (loosely) to a published test which was prepared by standard (or conventional) procedures. The usage of 'standardised' has become somewhat confused because published tests often present scores interpreted in terms of deviation from the mean (or average) and have a standard procedure for administering tests and interpreting results.

**Diagnostic test:** This term refers to the use made of the information gained from administration of the test. The implication is that the test results will assist in identifying both the topics which are not known and in providing information on potential sources of the student's difficulty. Teachers may be expected to provide appropriate teaching for each difficulty exposed by the use of a diagnostic test. For example, a simple open-ended mathematics question about area, given to junior secondary level classes provided a range of correct and incorrect answers. The question was:

A farmer built a fence around a rectangular plot of land. The longer sides were 5 metres and the shorter sides were 3 metres in length (See *diagram below*). What is the area of the fenced land?



Answers included 8 metres$^2$, 16 metres, 16 metres$^2$, 15 metres$^2$, 30 metres$^2$, and 225 metres$^2$. Those who gave 8 metres $^2$, 16 metres or 16 metres $^2$ as answers were confusing perimeter and area. They probably added 3, 5, 3, and 5 to obtain 16 or added 5 and 3 to obtain 8. Those who gave 30 metres$^2$ probably multiplied 5 by 3 twice and added the two results, while those who gave 225 metres$^2$ probably multiplied 3 by 5 by 3 by 5. Being able to show how the wrong answers were obtained may help the teacher to plan remediation (or the curriculum developer to devise suitable activities which will make the distinction between area and perimeter clearer so avoiding the problem). Some of those who gave the correct answer of 15 metres$^2$ showed their understanding of the task by sketching the 15 1-metre squares like this.




**Practical test:** In some senses an essay test is a practical task. The essay item requires a candidate to perform. This performance is intended to convey meaning in a practical sense by writing prose to an agreed format. However the term 'practical test' goes beyond performance and other tasks used in traditional pencil-and-paper examinations. The term may refer to practical tasks in trade subjects (such as woodwork, metalwork, shipbuilding, and leathercraft), in musical and dramatic performance, in skills such as swimming or gymnastics, or may refer to the skills required to carry out laboratory or field tasks in science, agriculture, geography, environmental health or physical education.

# 6 What types of task?

The kinds of question to be used in a test depend upon the age and learning experiences of the students, the achievements to be measured, the extent of the answer required, and the uses to be made of the information collected. The choice of tasks can also be influenced by the number of candidates and the time available between the collection of the evidence and the presentation of the results.

## Tasks requiring constructed responses

Some items require a response to be composed or constructed, whether written, drawn, or spoken. An essay question, for example, "Write three paragraphs describing the assessment context in your own nation and identify the key issues that need to be addressed", generally requires the candidate to compose several written sentences as the response. An oral test may have a similar task but the candidate is required to respond orally instead of in writing. The task may require production of a diagram, flow-chart, drawing, manipulation of equipment (as in finding the greatest mass using balance scales), or even construction (for example, weaving or building a model). More extensive tasks such as projects and investigations may require preparation of a report identifying the problem and describing the approach to the problem as well as the results obtained while attempting to solve the problem.

There are potential difficulties in scoring such prose, oral, drawn and manipulative responses. An expert judge is required because each response requires interpretation to be scored. Judges vary in their expertise, vary over time in the way they score responses (due to fatigue, difficulty in making an objective judgment without being influenced by the previous candidate's response, or by giving varying credit for some correct responses over other correct responses), and vary in the notice they take of handwriting, neatness, grammatical usage and spelling.

One technique to avoid or minimise such problems is to train a team of scorers. Such training often involves a discussion of what is being looked for, the key issues that have to be identified by a candidate. Then the scorers should apply what they have learned by scoring the same batch of anonymous real samples of responses. It is important to have a range of real samples. (The training is to ensure that scorers can tell the difference between high quality, medium quality, and low quality answers and assign marks so that the higher quality answers will get better scores than the medium quality answers, and medium quality answers in turn will get better scores than low quality answers.) These results are then compared (perhaps graphically) and discussed. The aim is not to get identical results for each scorer. Rather, the aim is to improve the agreement between scorers about the quality of each response. We expect that there should be greater agreement between the scorers where the responses are widely separated in quality. Making more subtle distinctions consistently requires more skill. Members of the scoring team may differ in the importance they place on various aspects of a task and fairness to all candidates requires consistency of assessment within each aspect. Even when team members agree in the rank ordering of responses, the marks awarded may differ because some team members are lenient while others are more stringent. A more subtle difference occurs when some judges see more 'shades of grey' or see fewer such gradations (as in the tendency to award full-marks or no marks).

Short-answer items may require a candidate to recall knowledge rather than recognise it (to produce an answer rather than make a choice of an answer) or may be restricted to recognition. The former may be something like miniature essays (or the oral or drawn equivalent), or may require a word or phrase to be inserted (as in close procedure or fill-the-gap). Recognition tasks may require a key element of a drawing/photograph/diagram/prose passage to be identified, as in the case of a proof-reading test of spelling or choosing the part of a diagram or poster which has a safety message.

Scoring short responses carries some of the same difficulties as scoring more extended responses but it is generally easier for judges to be consistent, if only because the amount of information to be considered is smaller and likely to be less complex. However, the quality assurance process is still a necessary part of the scoring arrangements for short responses. Tests that have only short responses may neglect the real world's need for extended responses.

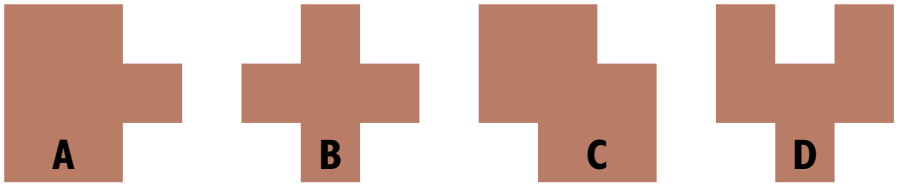
## Tasks requiring choice of a correct or best alternative

Some items present a task and provide alternative responses. The candidate's task is to identify the correct or the best alternative.

Sometimes such tasks require items in one list to be matched with items in another list but these tasks tend to be artificial; good tasks of this type are difficult to construct. Also, scoring may present problems when both lists are the same size. Those who are successful in choosing some of the links have their task of choosing the remaining links made easier. Those who are not successful with some links are faced with a more difficult task. It is not usually regarded as good practice to have success on one task influencing success on another separate task.

Some good matching items can be constructed if the number of links required in the answer is restricted. For example, this mathematics task requires only one link to be made out of a possible 6. (The six are A-B, A-C, A-D, B-C, B-D, and C-D. Note that the links can be written in reverse too: B-A, C-A, and so on.)

Two of these shapes have the same area.




Which two are they? ..... and .....

Multiple-choice items present some information followed by three or four responses, one of which is correct. The others, called distractors, are unequivocally incorrect, but this should be obvious only to candidates who 'know' that aspect of the work. An extreme case is where there are only two choices ( as in 'true-false', 'yes-no', "feature absent-feature present").

For example, the following multiple-choice item has four options, with only one the correct response.

The term *platyrrhini* refers to a group of animals which includes:

**A** Platypus

**B** Marmosets

**C** Flatworms

**D** Plankton

There are some potential difficulties with multiple-choice items. For example, it is possible to score this kind of test without knowing any answers to the items. The so-called 'correction for guessing' does not work – those who are lucky in guessing correct answers do not lose their advantage and those who are unlucky in their guessing do not get any compensation. Those that do not guess may be disadvantaged relative to those who have lucky guesses.

Further, using such a 'correction' increases the examiner's work and provides an opportunity for calculation errors which may reduce the accuracy of the scores.

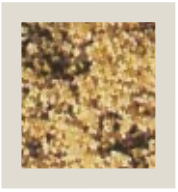
The probability of gaining high scores without knowledge is greater if there are only two choices. This factor, combined with the difficulty of constructing pairs of plausible choices, and the fact that 'correction for guessing' does not work, makes it unwise to use two-choice items (like 'true-false') in tests.

However, with a well-constructed test with an adequate number of items (each with three to five distractors), the probability of achieving a high score by random guessing is very small. If all items in a test are answered by all candidates, then applying a correction formula does not alter the rank order of candidates. In the educational context, most (if not all) tests should have sufficient time for most students to attempt most items. In this way, adequate time to attempt the items allows an adequate sample of performance to be gathered.

## Is one type of question better than another?

One important advantage of multiple-choice items is that the scoring is very consistent from marker to marker, relatively rapid, and can be undertaken by machine or by clerical staff. By contrast, performance tasks like essay items require markers skilled in assessing essays in the appropriate content area, take more time, and the markers may have problems in achieving consistency.

However, whatever type of question is used, the critical issue is whether the test provides a valid assessment of skills and knowledge in relation to the course objectives. It may be more appropriate to have items of *both* types in one test or examination (perhaps administered in separate sessions). It may also be necessary to combine test results with other evidence from practical tasks.



# 7 The test construction steps

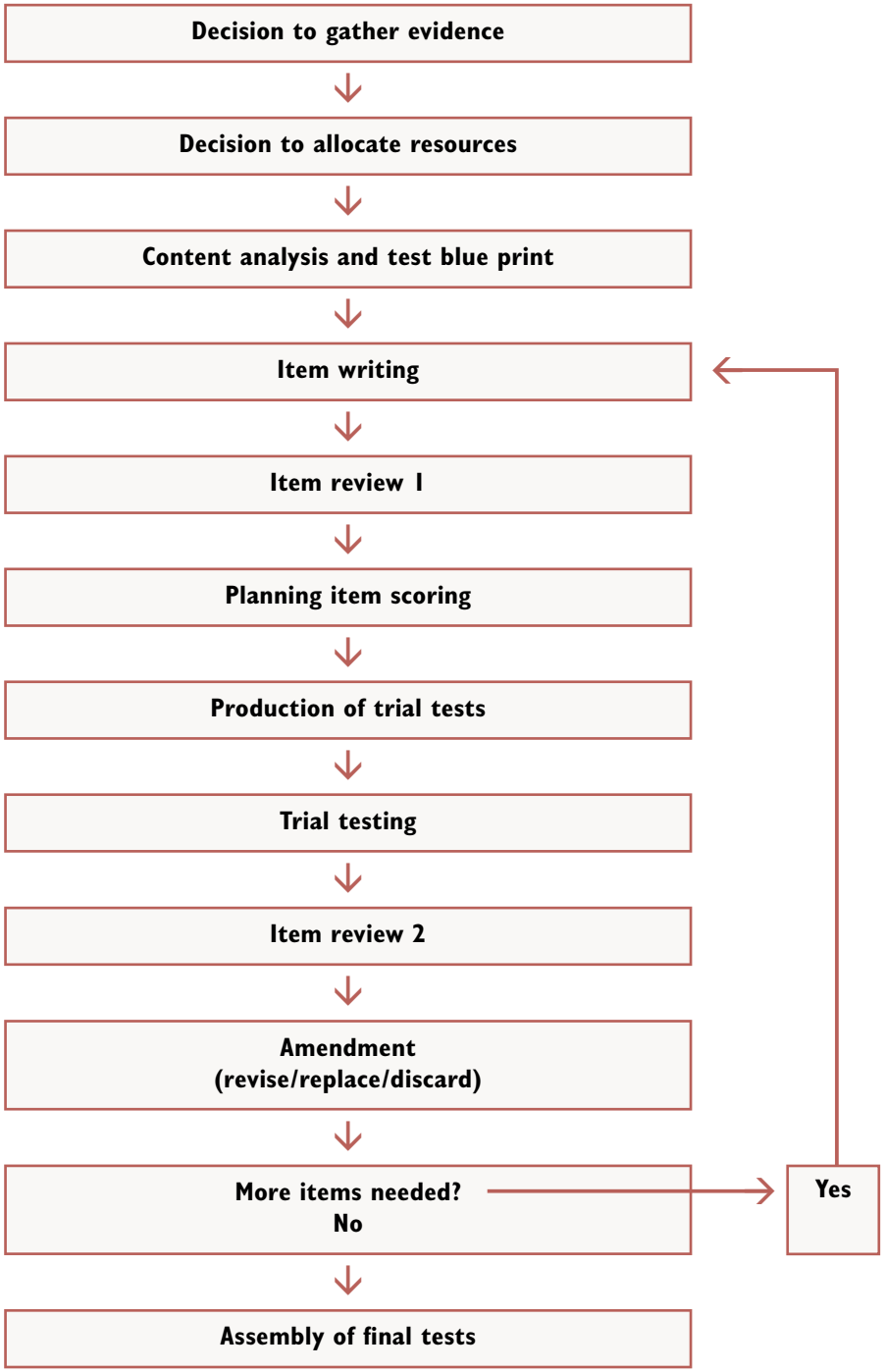
Before deciding to construct a test, one needs to know what information is required, how quickly it is needed, and the likely actions that are to be taken according to the results on a test. The crucial question is, «What information is needed about student achievement?» A second important question is, «Can we afford the resources needed to gather this information?» These resources include the costs involved in providing human resources, word processing and computing facilities, materials and equipment for photocopying and printing. *Figure 4* shows the stages in the test construction process.

## Content analysis and test blueprints

A content analysis provides a summary of the intentions of the curriculum expressed in content terms. Which content is supposed to be covered in the curriculum? Are there significant sections of this content? Are there significant sub-divisions within any of the sections? Which of these content areas should a representative test include?

A test blueprint is a specification of what the test should cover rather than a description of what the curriculum covers. A test blueprint should include the test title, the fundamental purpose of the test, the aspects of the curriculum covered by the test, an indication of the students for whom the test will be used, the types of task that will be used in the test (and how these tasks will fit in with other relevant evidence to be collected), the uses to be made

**Figure 4. Stages in test construction**



of the evidence provided by the test, the conditions under which the test will be given (time, place, who will administer the test, who will score the responses, how accuracy of scoring will be checked, whether students will be able to consult books (or use calculators) while attempting the test, and any precautions to ensure that the responses are only the work of the student attempting the test), and the balance of the questions. An example is shown in *Figure 5*.

This one-hour test is to assess prior knowledge of statistics of teacher trainees before they commence an intensive course in test construction and analysis. Items are to be multiple-choice in format with 4 options being presented for each item. A passing score is to be set for each content area; those below this cut-off score in an area must attend additional classes to improve their skills in that area. The 54-item test is to have several parallel forms and will be administered on a secure basis by the lecturer in charge of the test construction and analysis course. No books or calculators will be permitted for the test. Results will be provided the day after the testing. The test blueprint is shown in *Figure 5* below.

**Figure 5. Test blueprint for basic statistics**

| Content | Objectives | | | |
|---|---|---|---|---|
| | Recall of facts | Computational skills | Understanding | Total |
| Frequency distributions | 2 items | - | 4 items | 6 |
| Means | 2 items | 4 items | 2 items | 8 |
| Variances | 2 items | 4 items | 2 items | 8 |
| Correlation | 4 items | 4 items | 12 items | 20 |
| Relative standing | 4 items | - | 8 items | 12 |
| **Total** | **14** | **12** | **28** | **54** |

Comparing the test blueprint with the analysis of the curriculum should show that the allocation of items across the cells in *Figure 5* provides a reasonably representative sample of what the curriculum is about (at least as far as content is concerned). Test blueprints may include other dimensions too. For example, the blueprint may indicate the desired balance between factual recall questions and questions which require interpretation or application to a particular context. Or the blueprint may show the desired balance between different item formats (constructed responses as compared with recognition responses).

When the test blueprint has several dimensions it is possible to see how the evidence to be collected combines these dimensions with various types of evidence by means of a grid (or series of grids) and how account is taken of the importance of that evidence.

## Item writing

Item writing is the preparation of assessment tasks which can reveal the knowledge and skill of students when their responses to these tasks are inspected. Tasks which confuse, which do not engage the students, or which offend, always obscure important evidence by either failing to gather appropriate information or by distracting the student from the intended task. Sound assessment tasks will be those which students want to tackle, those which make clear what is required of the students, and those which provide evidence of the intellectual capabilities of the students. Remember, items are needed for each important aspect as reflected in the test specification. Some item writers fall into the trap of measuring what is easy to measure rather than what is important to measure. This enables superficial question quotas to be met but at the expense of validity – using questions that are easy to write rather than those which are important distorts the assessment process, and therefore conveys inappropriate information about the curriculum to students, teachers, and school communities.

# Item review

## The first form of item analysis: Checking intended against actual

Writing assessment tasks for use in tests requires skill. Sometimes the item seems clear to the person who wrote it but may not necessarily be clear to others. Before empirical trial, assessment tasks need to be reviewed by a review panel (with a number of people) with questions like:

- Is the task clear in each item? Is it likely that the person attempting an item will know what is expected?
- Are the items expressed in the simplest possible language?
- Is each item a fair item for assessment at this level of education?
- Is the wording appropriate to the level of education where the item will be used?
- Are there unintended clues to the correct answer?
- Is the format reasonably consistent so that students know what is required from item to item?
- Is there a single clearly correct (or best) answer for each item?
- Is the type of item appropriate to the information required?
- Are there statements in the items which are likely to offend?
- Is there content which reflects bias on gender, racial, or other grounds?
- Are the items representative of the behaviours to be assessed?

- Are there enough representative items to provide an adequate sample of the behaviours to be assessed?

This review before the items are tried should ensure that we avoid tasks which are expressed in language too complex for the idea being tested, avoid redundant words, multiple negatives, and distractors which are not plausible. The review should also identify items with no correct (or best) answer and items with multiple correct answers. Such items may be discarded or re-written.

## Other practical concerns in preparing the test

- How much time will students have to do the actual test? What time will be set aside to give instructions to those students attempting the test? Will the final number of items be too large for the test to be given in a single session? Will there be a break between testing sessions when there is more than one session?

- Will the students be told how the items are to be scored? Will they be told the relative importance of each item? Will they be given advice on how to do their best on the test?

- What test administration information will be given to those who are giving the trial test to students? Will the students be told that the results will be returned to them? Are the tests to be treated as secure tests (with no copies left behind in the venue where the test is administered)?

- Do students need advice on how they are to record their responses? If practice items are to be used for this purpose, what types of response should they cover? How many practice items will be necessary?

- Will the answers be recorded on a separate answer sheet (perhaps so that a test booklet can be used again)? Will this use of a separate sheet add to the time given for the trial test? What information should be requested in addition to the actual responses to the items? (This might include student name, school, year level, sex, age, etc.)

- Has the layout of the test (and answer sheet if appropriate) been arranged for efficient scoring of responses? Are distractors for multiple-choice tests shown as capital letters (easier to score than lower case letters)?

Have the options in multiple-choice items been arranged in some logical order (for example, from smallest to largest)? Have the items been placed in order from easiest to most difficult (to encourage candidates to continue through the test)? Has the layout of items avoided patterns in the correct answers such as 3 or more of the same letter in a row, or other patterns like ABCD or ABABAB (which might lead to 'correct' responses for the 'wrong' reasons)?

## Item scoring arrangements

**Multiple-choice:** Judgments of experts are needed to establish which option is the best (or correct) answer for each item. Once these correct answers have been decided, the score key can then be used by clerical staff or incorporated in machine scoring.

**Constructed response:** What preparation do the scorers need? Should they practice with a sample of papers to ensure that good work is given due credit, poor work is recognised consistently, and that each scorer makes use of similar ranges of the scale? Should each paper (or a sample of papers) be remarked without knowledge of the other assessment? If large differences occur in such a case, what should be the next step?

## Trial of the items

Item trial is sometimes called pilot testing – but in this context it does not mean testing those who fly aeroplanes. As well as considering the best efforts of item writers and item reviewers as a means of eliminating faulty items and improving the quality of items, it is necessary to subject the proposed items to empirical trial with students similar to those who are going to use the final form of the test. Since items involve communication with students, an evaluation of this quality is required before the set of tasks can be used with a larger group.

Each trial paper should be attempted by 150-250 persons who are similar to those who will attept the final forms of the test. It is usual to allocate the trial forms on a random basis within each trial examination room so that (on the average) each trial test is attempted by candidates of comparable ability. The same form of a test should *not* be given to candidates sitting in adjacent seats so as to ensure that candidates do not improve their scores by looking at another candidate's paper. It is wise to have some visible distinguishing mark on the front of each version of the test. Then the test supervisor can see at a glance that the trial tests have been alternated. If distinguishing marks cannot be used, then a different color of cover page should be used for each version.

Undertaking trial testing requires sound planning. Institutions which have agreed to allow trial testing to occur on their premises need to be contacted in advance of the trial testing. The numbers of trial candidates, the balance between males and females, the diversity of age levels or schooling levels required for the trials, the size of the rooms, and the availability of test supervisors are all issues that need to be discussed. The test supervisor introduces the test to the trial candidates, explains any practice items, and has to ensure that candidates have the correct amount of time allowed to attempt the test, that any last minute queries are answered (such

as informing those attempting the trials tests that their results will be used to validate the items and will not have any effect on their current course work), and gather all test materials before candidates leave the room.

## Processing test responses after trial testing

If the test needs to be scored before analysis, this scoring is done next. If there are essays to be scored, it is good practice to mark the first essay all the way through the stack of test papers. Then start the stack again to score the next essay. When all items have been marked, the scores on each item are entered into a computer file. If the test is multiple-choice in format, the responses may be entered into a computer file directly.

## Item analysis

### The second form involving responses by real candidates

Empirical trial can identify instances of confused meaning, alternative explanations not already considered by the test constructors, and (for multiple-choice questions) options which are popular amongst those lacking knowledge, and 'incorrect' options which are chosen for some reason by very able students.

This trial allows the gathering of evidence about each item – whether items can distinguish those students who are knowledgeable from those lacking knowledge, whether items are of an appropriate difficulty (how many attempted each item and what percentage responded correctly), and, in the case of multiple-

choice items, whether the various options, both 'correct' and 'incorrect' performed as expected. The item analysis also provides an opportunity to collect information about how each item performs relative to other items in the same test, and to judge the consistency of the whole test.

## Amending the test
### by discarding/revising/replacing items

Items which do not perform as expected can be discarded or revised. However discarding questions when there is a shortage of replacement questions can lead to distortions of the achieved test specification. If the original specification represents the best sampling of content, skills, and item formats, in the judgments of those preparing and reviewing the test, then leaving some cells of the grid vacant will indicate a less than adequate test. To avoid this possibility, test constructors may prepare three or four times as many questions that they think they will need for *each cell* in the grid.

## Assembling the final test
### (or a further trial test) and the corresponding score key

After trial, tasks may be re-ordered to take account of their difficulty. Usually the easiest questions are presented first. This is to encourage candidates to proceed through the test and to ensure that the weaker candidates do not become discouraged before providing adequate evidence of their achievements and skills. Minor changes to items may have to be made for layout reasons (for example, to keep all of an item on one page of the test, or to avoid obvious

patterns in the list of correct answers). Items representing a single cell within a test specification should vary in item content and difficulty. The position of the correct option in multiple-choice items (A, B, C, D or E) should also vary and each position should be used to a similar extent. Some questions may have minor changes to wording, others may be replaced. The final test should be consistent with the test blueprint. The item review procedures described above are repeated (particularly important where stimulus material must be associated with more than one question) and each reviewer should work independently through the proposed test and provide a 'correct' answer for each question. This enables the test constructor's (new) list of correct answers to be checked.

## Validity, reliability and dimensionality

Test validity can be interpreted as usefulness for the purpose. Since purposes vary, it is important to specify which purpose applies when making a comment about validity. Content validity refers to the extent to which the test reflects the content represented in curriculum statements (and the skills implied by that content). A test with high content validity would provide a close match with the intentions of the curriculum, as judged by curriculum experts and teachers.

A test with high content validity for one curriculum may not be as valid for another curriculum. This is an issue which bedevils international comparisons where the same test is administered in several countries. Interpretation of the results in each country has to take account of the extent to which the comparison test is content valid for each country. If two countries have curricula that are only partly represented in the test, then comparisons between the results of those countries are only valid for part of the data.

When test results are compared with an agreed external criterion such as a direct measure of actual performance of tasks in the 'real' world, this type of validity is called criterion-related validity. If there is little time delay between the test and the actual performance, the criterion-related validity may be referred to as concurrent validity. If there is a longer time delay between the test and subsequent actual performance, the criterion-related validity may be referred to as predictive validity.

If we think in terms of achievement as a generalized construct, and our test tends to be consistent with other recognized measures of that construct, we say that the test has construct validity as a measure of achievement. Similarly, if we think in terms of aptitude as a generalized construct, and our test tends to be consistent with other recognized measures of that construct, we say that the test has construct validity as a measure of aptitude. The higher the degree of agreement, the higher the construct validity. However, this is not a fixed state of affairs. Particular tests may have high construct validity as achievement measures or predictive validity as an indicator of later success. If circumstances change (such as teachers teaching to that particular test or tests) the scores on the test may well rise considerably without improving the predictions. The assumed association between the test and the predicted behaviour no longer holds, and raising the cut-off on the test will not rectify the problem.

When tests have high construct validity we may argue that this is evidence of dimensionality. When we add scores on different parts of a test to give a score on the whole test, we are assuming dimensionality without checking whether our assumption is justified. Similarly, when item analysis is done using the total score on the same test as the criterion, we are assuming that the test as a whole is measuring a single dimension or construct, and the analysis seeks to identify items which contradict this assumption.

Earlier in this discussion, it was argued that validity refers to usefulness for a specified purpose and can only be interpreted in relation to that purpose. In contrast, reliability refers to the consistency of measurement regardless of what is measured. Clearly, if a test is valid for a purpose it must also be reliable (otherwise it would not satisfy the usefulness criterion). But a test can be reliable (consistent) without meeting its intended purpose. That is, it is possible to make consistent assessments which are not meaningful in the context of the decisions to be made.

# 8 Resources required to construct and produce a test

Many teachers have some skills in preparing assessment tasks but receive little feedback on which tasks are valid and useful. Those preparing questions for national examinations may receive some feedback on the quality of the assessment tasks they have prepared, but only if the examining authority conducts the appropriate analyses. Without such quality feedback the skill level of item writers tends to remain low. Expertise in developing non pencil-and-paper assessment tasks is an even more scarce resource.

Training in test construction requires an expertise-sharing approach so that the test construction skills of the trainer are transferred to those involved in assessment. The team writing the questions has to be aware of actual candidate responses to those questions, and to have an opportunity to discuss the subsequent analyses of trial data. Development of expertise is incremental, requiring an ability to distinguish between what was intended and what actually happened in practice. As experience in developing tests, administering them, and interpreting the resulting responses is gained, there should be less involvement with external trainers (and more involvement in sharing that developed expertise with novice test constructors).

Producing a test has a number of costs, aside from the physical provision of weather- and vermin-proof secure room space, heating or cooling, furniture and secure storage. The costs relate to developing a test specification, the test construction effort by

teachers (either set aside from classroom work to join the test-construction team, or paid additional fees to work outside school hours), class teacher and student time for trials, the paper on which the test (and answer sheet if appropriate) is to be printed or photocopied, the production of copies of the test materials, distribution to schools, retrieval from schools (if teachers are not to score the tests), and scoring and analysis costs. *Figure 6* shows a possible time scale for developing two parallel forms of an achievement test of 50 items for use during the sixth year of schooling. *Figure 6* also shows the resources that would need to be assembled to ensure that the tests were produced on time. (Note that this schedule assumes that the test construction team has had test development training prior to commencing work on the project.)

## Figure 6. Example timescale and resource requirements for test construction

| Task | Time (weeks) | Resources |
|---|---|---|
| Decision to gather evidence | 1 | (Depends on local circumstances) |
| Decision to allocate resources | 2 | (Depends on local circumstances) |
| Content analysis and test blueprint | 1 | Curriculum experts (national or regional).<br>Test construction team (see below).<br>Relevant text books as used in the sixth year of schooling. |
| Item writing | 5 | Test construction team (3 to 4 teachers, full-time).<br>Text books, word-processing, and photocopying facilities and supplies. |
| Item review 1 | 1 | Curriculum experts & test construction team. |
| Planning item scoring | 1 | Test construction team. |
| Production of trial tests | 2 | Test construction team.<br>Word-processing, and photocopying facilities and supplies. |
| Scoring and item analysis | 3 | Test construction team.<br>PC for computing. |
| Item review 2 | 1 | Curriculum experts & test construction team. |
| Amendment (revise/replace/discard) | 1 | Word-processing, and photocopying facilities and supplies. |
| More items needed? *No*/*Yes* | | If '**yes**' *Go back to* 'Item writing'.<br>If '**no**' continue. |
| Assembly of final tests | 2 | Test construction team.<br>Word-processing, and photocopying facilities and supplies. |
| **Total Time** | **20 weeks** | |

# 9 Some concluding comments

At the beginning of this module it was asserted that the assessment of student learning provides evidence so that sound educational decisions can be made. This evidence should help us to evaluate (or judge the merit of) a teaching programme or we may use the evidence to make statements about student competence or to make decisions about the next aspect of teaching for particular students. Clearly the quality of the evidence is a critical factor in making sensible decisions.

The procedures for test construction described in this module have developed over many years of practical work in the development of tests and similar instruments. Some of the advice has arisen from research into test analysis and some advice has been derived from the practical experience of large numbers of research and development staff working at various agencies around the world. Improving the quality of the evidence is not an easy task. And reading a book about the procedures will not suffice – because improving one's skills as a test constructor requires working on the construction of tests as part of a test construction team.

Preparation of final forms of a test is not the end of the work. The data gathered from the use of final versions should be monitored as a quality control check on their performance. Such analyses can also be used to fix a standard by which the performance of future candidates may be compared. It is important to do this as candidates in one year may vary in quality from those in another year.

It is customary to develop more trial forms so that some forms of the final test can be retired from use (where there is a possibility of candidates having prior knowledge of the items through continued use of the same test).

The trial forms should include acceptable items from the original trials (not necessarily items which were used on the final forms but similar in design to the pattern of item types used in the final forms) to serve as a link between the new items and the old items. The process of linking tests using such items is referred to as *anchoring*. Surplus items can be retained for future use in similar test papers.

# 10 References

## General measurement and evaluation


1. Hopkins, C.D. and Antes, R.L. (1990). *Classroom measurement and evaluation.* Itasca, Illinois: Peacock.

2. Izard, J. (1991). *Assessment of learning in the classroom.* Geelong, Vic.: Deakin University.

3. Mehrens, W.A. and Lehmann, I.J. (1984). *Measurement and evaluation in education and psychology.* (3rd Ed.) New York: Holt, Rinehart and Winston.


## Content analysis and test blueprints


1. Izard, J. (1997). *Content Analysis and Test Blueprints.* Paris: International Institute for Educational Planning.


## Item writing


1. Withers, G. (1997). *Item Writing for Tests and Examinations.* Paris: International Institute for Educational Planning.


## Trial testing and item analysis


1. Izard, J. (1997). *Trial Testing and Item Analysis in Test Construction.* Paris: International Institute for Educational Planning.

## Testing applications


1. Adams, R.J., Doig, B.A. & Rosier, M.J. (1991). *Science learning in Victorian schools: 1990.* (ACER Research Monograph No. 41). Hawthorn, Vic.: Australian Council for Educational Research.

2. Doig, B.A., Piper, K., Mellor, S. & Masters, G. (1994). *Conceptual understanding in social education.* (ACER Research Monograph No. 45). Melbourne, Vic.: Australian Council for Educational Research.

3. Masters, G.N. et al. (1990). *Profiles of learning: The basic skills testing program in New South Wales, 1989.* Hawthorn, Vic.: Australian Council for Educational Research.

4. Ross, K.N. (1993). *Issues and methodologies in educational development: 8. Sample design procedures for a national survey of primary schools in Zimbabwe.* (International Institute for Educational Planning) Paris, France: United Nations Educational, Scientific and Cultural Organisation.


## Information for decision-making


1. Somerset, A. and Ekholm, M. (1990). "Different information requirements for different levels of decision-making", in K.N. Ross and L. Mählck, L. (eds.) (1990). *Planning the quality of education: The collection and use of data for informed decision-making.* Paris: United Nations Educational, Scientific and Cultural Organization/Oxford: Pergamon Press.

# 11 Exercises

## 1. CONSTRUCTION OF A TEST PLAN

**a)** Choose an important curriculum topic or teaching subject (either because you know a lot about it or because it is important in your country's education programme).

List the key content areas in that topic or subject.

Show (in percentage terms) the relative importance of each key area.

Compare your key content areas and associated relative importance with one or more persons attempting this exercise.

**b)** Choose another appropriate dimension (such as skills categories or item format categories) for the same curriculum topic or subject (as in *Exercise 1* above).

List the important categories.

Show the relative importance (expressed as percentages) of each category.

Compare your categories with one or more persons (as in *Exercise 1*).

**c)** Construct a test plan (like the plan shown in *Figure 5*) which has the content categories (from *Exercise 1*) at the left and the skill or format categories (from *Exercise 2*) at the top.

Adjust the numbers of items in each cell to reflect the percentage weightings you have chosen for each dimension.

## 2. *TEXTBOOK ANALYSIS AND ITEM WRITING*

**(a)** Review a classroom textbook used in your country. Using your test plan as a guide, prepare a test plan for a test of the material in the text book.

**(b)** Choose one cell of the test plan and write some items for this cell.

## 3. *REVIEW OF EXISTING CLASSROOM TEST*

**(a)** Using the section on item review as a guide, review a classroom test prepared by a teacher.

**(b)** Set up a panel of two or three to discuss the reviews.

**(c)** Choose some of the questions and re-write them to satisfy the panel's critical comments.