

1

Unit 1: Defining the Assessment Needs

Unit 1 of the CCET (Level A) deals with a number of key issues in testing. Assesseees are expected to be aware of the nature of the relationship between testing and assessment and be able to identify assessment needs that would best be dealt with by using a test (and conversely, those for which a test would not be appropriate). They are also expected to be able to justify the choice of a particular test with reference to the learner's development and the skills being assessed, and to show how the results from tests of ability and attainment can be affected by factors such as ethnicity, gender, social class and disability, as well as more direct factors in the testing environment, such as heat, noise, test anxiety and variables associated with test administration. Assesseees should also be able to identify sources of information that can corroborate the information from a test, and be able to distinguish between different types of test. We shall consider each of these issues in turn.

The Relationship between Testing and Assessment

There will be few, if any, educational test users who are unfamiliar with the process of assessment, which is defined by Salvia and Ysseldyke (2004, p. 4) as: 'a process of collecting data for the purpose of making decisions about individuals and groups'.

This identifies three aspects of assessment:

- It involves a *process over time*, not just the observation of how a learner performs on one occasion.

- It is always carried out for a *reason*, for example to answer particular questions such as ‘*How do the learner’s reading skills compare with those of other pupils of the same age?*’ or ‘*What are the learner’s strengths and weaknesses in regard to learning new phonic skills in reading?*’
- It should test one or more specific *hypotheses* about the learner’s performance (for example, about possible factors that may precipitate or maintain a problem or explain why a problem may have occurred), which will both increase understanding and provide support for decisions as to how best to proceed.

Real-life problems are complex with many inter-related dimensions. For example, a difficulty in reading may result in problems with self-esteem and motivation, with resultant reluctance to complete classwork and homework, possible problems in peer relationships with classmates or behaviour difficulties in class, and tensions between home and school. The process of assessment has to reflect this complexity by gathering, structuring and analysing information to identify the key dimensions of the problem and understand how they interact, so that a clearer understanding both of the problem(s) and of what must be done to improve the situation can be generated.

To achieve this, information may be gathered by education professionals in collaboration with stakeholders, such as parents/carers, the students/pupils/learners themselves, and any others who may be involved, to develop a shared understanding of the nature of the problem and of the intervention required to assist decision making and achieve change. This information can be gathered in respect of an individual, or a group, for example, a class, and is a key component of the process of teaching and learning.

Such information is commonly collected in respect of concerns about academic performance, generally on account of concern about progress in the curriculum or some aspect of basic skills in language, literacy or numeracy; problems in behaviour, such as disruption, difficulties in social relationships, or withdrawn, shy behaviour; or the effects of known physical problems, for example, with vision, hearing, motor skills or mobility, or traumatic head injury.

But as Salvia and Ysseldyke (2004, pp. 23–38) and Urbina (2004, pp. 23–6) note, it is important to ensure that we gather information *from a range of sources*, (using both available information and cumulative information, such as portfolios of evidence of the learner’s work) and *in different contexts* (to make the process as representative as possible) about:

- The *teaching and learning environment*, for example, whether the learner’s teaching arrangements and the teaching techniques and

curriculum employed are appropriate for his/her needs in regard to interest and relevance, level of difficulty, progression, presentation, and the nature of the response required; the social context of the setting and relationships with other learners; and the organization of the settings in which the learner's teaching and learning takes place.

- Factors outside the immediate educational setting, from *home and family, school and community*, which affect the learner's performance, for example, the experience he/she brings to school from home; the effects of whole-school factors, such as ethos, policies, legislation, guidelines and school development plans; and community influences upon attitudes towards learning and achievement).
- The *learner*, for example, his/her levels of confidence and independence as a learner; motivation; knowledge and use of strategies; levels of existing skills, knowledge and understanding in regard to a task or activity; learning styles; the extent to which he/she can generalize from what has been learned by reflecting upon the activity and seeing its relevance and connections to existing knowledge; and effects and influences of gender, ethnicity and social class.

There are a number of ways in which we can do this, including:

- *Direct observation* of the student/pupil/learner's performance or behaviour, perhaps by means of a systematic approach such as an observation schedule to investigate specific pre-defined behaviours, such as time spent 'on task', or a less systematic approach, such as a 'running record' of everything the learner does during a specified period of time. However, as well as requiring a considerable amount of time, observations can be subject to bias resulting from subjective judgements on the basis of labels (for example, such as 'gifted' or 'dyslexic'), physical attributes, other characteristics of the learner, or from *expectancy effects* (Salvia and Ysseldyke, 2004, pp. 27–8), where our beliefs colour what we observe, and 'we see what we expect to'. Thus, a teacher carrying out an evaluation of a behaviour management programme in his/her class may be more likely to observe negative behaviours if they have no confidence in the programme, and in contrast, more likely to observe positives if they believe that the programme has been successful.
- *Existing information*, such as portfolios of the learner's work, which can be compared with those of others in the class, can also be used, although interpretation will be restricted to what information is available and to the circumstances in which it was gathered.
- *Interviews* with those who know the learner well, such as former teachers or parents, also provide useful information.

- *Professional judgements* of those who have worked with the learner in the past or are currently working with the learner in the here and now. However, sources of bias include the possibility that key information may be forgotten, that the informant may not recognize the importance of particular information, or that informants may be selective with the information they provide.
- Finally, *a test* may be used.

Test use is a sub-set of the assessment process (Salvia and Ysseldyke, 2004, p. 6). But what do we mean by a ‘test’? One of the leading experts in test use, Lee J. Cronbach (1984, p. 26), noted that ‘There is no fully satisfactory definition for *test*’. With this caveat, let us consider some attempts at a definition:

A test is a predetermined set of questions or tasks for which predetermined types of behavioral responses are sought. Tests are particularly useful because they permit tasks and questions to be presented in exactly the same way to each person tested. (Salvia and Ysseldyke, 2004, p. 29)

A psychological test is a measurement instrument that has three defining characteristics:

- A psychological test is a sample of behaviour.
- The sample is obtained under standardised conditions.
- There are established rules for scoring, or for obtaining quantitative (numeric) information from the behaviour sample. (Murphy and Davidshofer, 1994, p. 3)

A psychological test is a systematic procedure for obtaining samples of behaviour relevant to cognitive or affective functioning and for scoring and evaluating these samples according to standards. (Urbina, 2004, p. 1)

From these definitions, we can see that a test is a form of systematic assessment, with standardized procedures, from which numerical scores are taken. But as Urbina (2004, p. 25) notes, use of a test differs from other assessment approaches in the following ways:

- Testing is quicker and simpler than a more general assessment approach involving a range of processes, such as interviewing or observation.
- In testing, the individual test taker is the principal source of data, whereas there are other sources of information, for example, teachers, family, peers routinely involved in a more general assessment approach.

- Testing tends to compare the test taker(s) with others, whereas assessment may also consider the uniqueness of the individual or group or the context.
- Testing is highly structured, with objective procedures, whereas assessment may be less structured and more subjective, relying more on professional judgements.

But it is worth recalling that tests not only provide quantitative data, that is, the actual scores obtained, but also qualitative data in the form of observations about the test taker's performance or how the performance was achieved. As we shall see in Unit 8, such qualitative data can be important in the interpretation of the results from a test.

There are many different types of test. Rust and Golombok (2000, pp. 38–47), for example, identify the following:

- *Norm-referenced* tests (where an individual's scores are compared with those for a representative sample from a population) versus *criterion-referenced tests* (where scores are related to an external criterion, usually the completion of a task to a predefined standard).
- *Knowledge-based* tests (measuring ability, attainment, achievement and aptitude) versus *person-based* tests (measuring attributes such as personality, attitudes, moods, self-esteem, and social competence using rating scales or other self-report measures). Knowledge-based tests are tests of *maximum performance*, where the score is determined by the test taker's success in completing a well-defined task, that is, a measure of what the test taker can do in the test situation. Most standardized tests used in education, such as measures of ability, attainment or achievement, are of this kind. Person-based tests are tests of *typical performance*, where there is no 'right' or 'wrong' answer.
- *Objective* tests (with pre-specified scoring criteria) versus *open-ended* tests, such as essays or projective tests, where ambiguous pictures are presented and the test taker asked to describe what they see.

Salvia and Ysseldyke (2004, p. 29) identify two related dimensions that we need to take account of in making sense of test scores: *performance standards* and *informational context*. Performance standards provide the basis for comparisons between the learner and other students/pupils (a comparative standard) or between the learner's scores from two administrations of the same test (an absolute standard). But the learner's performance must also be contextualized and related to other information gathered. Good assessment – and test use – in this way utilizes *multiple sources of*

information in the interpretation of results to assist the decision-making process.

So when should a test be used? When you want to compare the results with those from other students/pupil/learners. As we have noted, it is important to factor in contextual and wider environmental variables, such as teaching approaches or the curriculum, or the learner's interests or views, when carrying out an assessment. But this can be a drawback if you want to draw comparisons across teachers or schools. Tests are designed to be presented in a standardized way and to elicit performances that can be scored in a consistent and comparable fashion, irrespective of the individual who administers the instrument, provided that he/she is trained and competent to do so, and adheres strictly to the procedures for administration and scoring.

Identifying Assessment Needs Which Would Best be Dealt With by Using a Test (and Conversely, Those For which A Test Would not be Appropriate)

The key to effective assessment is to match the strategy for gathering information to the questions being asked. For example, a teacher who has concerns about a pupil's progress in mathematics and asks a SENCO colleague for ideas to help tailor instruction to the pupil's areas of difficulty will be underwhelmed by assessment findings which merely confirm that the pupil has problems in mathematics.

Good practice consists of two steps: first, clarifying the purpose of assessment and then second, selecting the most appropriate, reliable and valid means of gathering information (which we will consider in more detail in Unit 3). Test use can be highly informative, but users need to be aware of the 'pros' and 'cons' of different approaches to assessment so that they can select the most appropriate means of answering questions.

The CCET (Level A) requires test users in education to be familiar with the following four approaches to assessment:

- 1 *Norm-referenced* assessment: where we compare the learner's performance with that of the other individuals in the normative sample of the test on whom the test norms are based, using *developmental scores* (for example, age-equivalent scores, such as 'Reading Ages' or 'Spelling Ages') or *derived scores* (for example, standard scores, or percentiles). Use a norm-referenced test when you wish to compare the pupil's knowledge base with norms (i.e. normative standards for his/her age).

- 2 *Criterion-referenced* assessment: where we compare the learner's performance against a specific, objective, pre-defined standard, or *criterion* on measures of single skills (for example, 'Can Jacqui correctly add $3 + 4$?) or on measures of more complex multiple skills (for example, 'Can Nicky read aloud pages 3 to 6 of his reading book with 90 per cent accuracy?'). Use criterion-referenced tests when you wish to assess the content of a pupil's knowledge base within a specific domain (i.e., when you want to measure mastery of information or skills: what the pupil can or can't do in terms of absolute standards).
- 3 *Curriculum-based* assessment: a derivative of the above, where we compare the learner's performance with the content of his/her existing curriculum. Use a curriculum-based test when you wish to assess the pupil's instructional needs based upon his/her on-going performance on the existing content of curriculum.
- 4 *Dynamic* assessment: an approach which holds that a learner's performance will improve with appropriate support and follows a 'test – teach – re-test' sequence, providing training in the form of guidance using cues and graded prompts. This approach highlights the importance of assistance (referred to as *mediation*, *facilitation* or *scaffolding*) and of the learner's responsivity towards tasks. Use this approach when you wish to find out more about the appropriateness of the strategies used by the student/pupil/learner, how they react to suggestions made by a teacher, and identify barriers to learning. Proponents of dynamic assessment view norm-referenced, criterion-referenced and curriculum-based approaches as 'static', focusing upon outcomes in achievement with little attention paid to the processes which underpin learning. There are published dynamic assessment 'tests' (see Tzuriel, 1990; 2000), with an emphasis upon mediation.

But as Rust and Golombok (2000, p. 42) point out, the items of all tests have to be related to some criteria, which are determined by the purpose of the test. Consequently, all tests, even those which are norm-referenced may be regarded as being criterion referenced.

Assessment needs in education are of four main types:

- 1 Those related to *entitlement* or *eligibility* for additional support:
 - Does a pupil have a significantly more marked, learning difficulty than the majority of learners of his/her age which makes him/her eligible for additional support (for example, specialist teaching or special examination arrangements on account of dyslexia)? This may involve individual assessment or perhaps screening a class to identify learners with additional support needs.

- Should a learner be referred to a specialist agency?
 - Documenting the nature and range of a learner's additional support needs.
- 2 Those related to *accountability*:
- Evaluating progress of a support plan or programme;
 - Evaluating outcomes from special initiatives;
 - Determining 'value added'; and
 - Monitoring standards of attainment in literacy and numeracy in a class.
- 3 Those related to *instruction*:
- Identifying what pupils have learned to inform the planning of what they should be taught;
 - Identifying strategies to improve pupils' attainments;
 - Deciding how to group pupils; and
 - Testing hypotheses about the nature of a pupil's additional support needs to help in the formulation of a programme.
- 4 Those related to *gathering information over time from a range of specific contexts*:
- Determining the nature and extent of a pupil's behavioural difficulties (which are often context-specific) by classroom observation;
 - Monitoring 'on-task' behaviour in class; and
 - Evaluating the effectiveness of intervention for pupils with social, emotional or behavioural difficulties.

Assessment need	Possible assessment method
Determining entitlement or eligibility for additional support	Norm-referenced, standardized tests
Establishing accountability	Norm-referenced, standardized tests
Gathering information about instructional needs	Criterion-referenced, curriculum-based or dynamic assessment approaches
Gathering information about situated, or context-specific problems, such as social, emotional or behavioural difficulties	Interviews and observations in different settings over a period of time

As the table above shows, there is a degree of overlap between assessment approaches and assessment needs. However, norm-referenced (*psychometric*), standardized tests are particularly informative in addressing assessment needs related to entitlement/eligibility and accountability issues, but

somewhat less so in regard to issues regarding instruction or the details of intervention, although the scores and associated qualitative data could yield relevant information. But in general, criterion-referenced, curriculum-based or dynamic assessment approaches would be more appropriate in regard to addressing assessment needs relating to instruction. As Lidz (1991, pp. 121–2) puts it:

If we wish to determine how far the child's knowledge base deviates from the norm, we will continue to administer a psychometric measure. If we wish to determine the content of a child's knowledge base within a specific domain, we will administer a curriculum-based test. If we wish to derive hypotheses about how the child learns, how responsive the child is to attempts to intervene, and what seems to be interfering with the child's ability to profit from existing attempts at instruction, we will use dynamic assessment.

However, there are other assessment needs, for example, associated with situated, or context-specific problems, such as social, emotional or behavioural difficulties, where test use would be less helpful. With such problems, interviews gathering information about the history and nature of the concerns, and observations in different settings over a period of time would yield information about the difficulties themselves and their history, as well as the characteristics of the individual and relevant factors in the contexts in which the problems are presented and would be helpful in suggesting how best to proceed.

Justifying the Choice of a Particular Test with Reference to Knowledge of the Learner's Level of Development and the Skills being Assessed

One possible pitfall in test use is to fall into the habit of using the same test or tests over and over again without adequate consideration of the purpose of carrying out the assessment in the first place. One rule of thumb to help guard against this is to ask the question: 'What will I do differently once I have the test results?' If you do not have an answer to this, then the chances are that you are not clear about why you are using the test.

Competent test users should be familiar with any test that they are anticipating using. They need to know the details of the materials and must be practised in administration. But equally, they should be able to provide a rationale and justification for the selection of a particular test taking into the account not only the purpose of assessment, as discussed

above, but also the learner's level of development and the skills being assessed.

Learner's level of development

In the case of the learner's level of development, this means taking factors such as the following into account:

- The *appropriacy of the norms* (for example, the representativeness and size of the norm group, the age-range covered, and whether the norms are up-to-date, in the case of standardized tests). The learner's scores on a standardized test will be compared with the scores of the participants in the norm group. Accordingly, the test user must ensure: (i) that the norms of any test used are based upon participants who are representative of the test takers; (ii) that they are based upon an adequate number of participants (see Unit 2 for discussion); (iii) that the age-range of the test is appropriate for the test taker; and (iv) that the date of standardization is no more than 10 years from the date of administration for achievement tests and no more than 15 years for ability tests (see Salvia and Ysseldyke, 2004, pp. 114–15 for a discussion).
- The *appropriacy of the level of the items*. If the test norms are to be used to score the test taker's performance, the test user must ensure that the learner is not administered items which are outside of the age-range for his/her age. If out-of-age level items are used in this way, the test taker is not being compared with same-age participants and the score would be invalid. But in cases where the learner has marked difficulties, the test user should also consider whether the materials are appropriate for the test taker's age and developmental level, even if the items have been standardized on an appropriate age-group.
- The *interest level of the test items*. Care must be taken to ensure that the test taker does not consider the materials to be too difficult or too 'easy'. This may be a particular consideration in the case of a younger learner who may lack self-confidence and whose interest is not engaged, or in the case of an older test taker with learning difficulties, who may consider that the materials are more appropriate for a younger learner.

The skills being assessed

In the case of the skills being assessed, this means taking into account factors that relate to the accessibility of content of materials:

- The *accessibility of the test materials*. The test medium (for example, the format of the test, the mode of administration and the mode of

presentation, basal and ceiling levels) and the language of the test items (for example, the vocabulary used, any literacy demands that might be made and any cultural referents) are important considerations when selecting a test. If a test taker cannot understand what he/she is required to do, then that measure is clearly invalid as a test of the particular skill or ability which it is intended to assess. This may even contravene equal opportunity and/or disability legislation. For example, tests presented orally to test takers with severe hearing impairments unable to understand the instructions and task demands would be unfair. Such instruments could also be problematic for those with severe language and communication problems or for those for whom English is an additional language if not presented in the test taker's primary language. Similarly, test items with instructions presented in print form would be unfair not only for test takers with marked visual impairment, but also for those with problems in reading, who are unable to access the meaning of what they are required to do. With regard to cultural relevance and referents, the test materials should not give an unfair advantage to test takers from a particular religion, ethnic group or social class, nor offence to others. The choice of selecting an individually administered or a group-administered test may also have an effect upon the test taker's performance. For example, group-administration of a test may have significant sources of distraction, which could disadvantage a learner with poor concentration or attention problems. It may also be harder to ensure that all of the test takers are clear about what they are required to do. Finally, a test should have clear instructions for establishing a *basal level*, the level below which it is assumed that the test taker will get all of the items correct, and a *ceiling level*, the point at which it is assumed that the test taker will fail all items and testing is discontinued.

- *The appropriacy of the response to the test material.* The test format will determine how the test taker is required to respond to the test items. This may be: (i) verbally; (ii) in writing, as in the case of a multiple-choice test or rating scale; or (iii) perhaps by means of a motor response, as in some non-verbal performance tests which may involve the manipulation of blocks or other objects. But if a test taker's sensory or physical difficulties were to constrain his/her ability to respond, then the test results would be invalid, and perhaps also illegal. Thus a learner with a severe speech and language delay or disorder may have difficulty in providing a clear verbal response, and a learner with a physical disability or motor difficulty, such as developmental coordination disorder/dyspraxia, may have problems responding in writing or via the manipulation of objects, particularly if the test has time limits. Again, the selection of a group-administered rather than an individually

administered test would also be a consideration as it is likely that there would be more reliance upon writing skills, and possibly reading skills in the case of a group-administered test.

- *The learner's exposure to the material being tested.* A test taker who has not had experience of the material being tested, or perhaps not as much exposure to the curriculum being assessed as his/her peers, will be at a relative disadvantage to a test taker who has. Thus, a learner with poor school attendance, or a test taker recently re-located from a different country or culture would be at a disadvantage in regard to a test of a curricular area if he/she missed the opportunities for instruction afforded to good attenders. In such cases, the results from the test might reflect more the lack of opportunity to learn than problems in the skill or ability being tested and would raise concerns about the fairness of the test use.

The Effects of Factors such as Ethnicity, Gender, Social Class and Disability, as well as More Direct Factors in the Testing Environment, such as Heat, Noise, Test Anxiety and Variables Associated with Test Administration upon the Results from Tests

Effective test use depends not only upon careful selection of the instrument and practice in administration and scoring on the part of the test user, but also in ensuring that possible sources of bias and distraction in the testing environment are removed or at least minimized. This is particularly important where test takers are children, indeed children with additional support needs. Known environmental variables that have a direct and adverse effect upon the test taker's concentration and ability to attend and hence his/her performance include heat, poor lighting, noise and over-crowding. The test user should thus ensure:

- that the session is *free from interruptions* (a 'Do not disturb' sign outside the test room may be helpful); and
- that the room used for testing is *adequately ventilated*, with an appropriate background temperature; *well-lit*, *free from extraneous noises* and where possible free from other possible distracters (such as food or drink, or third parties who might influence the test taker's response); and has *appropriate seating and space* for test takers (for example, a desk and chair of the correct height, and well-spaced out desks or tables for administration of a group test to minimize copying of answers).

In addition, there are cognitive and emotional variables that are important. Tiredness and lack of motivation, for example, can adversely affect the child's performance, as can *test anxiety*, where the experience of being evaluated, fear of failure, the context in which testing takes place or even the specific type of test used (for example, a timed test or perhaps a test of mathematics) causes the test taker to feel anxious. While a degree of such anxiety may increase the test taker's physiological levels of arousal and actually improve their performance, an excess can have a catastrophic effect.

Variables associated with the *test administration* can also have an effect upon the results. If the test user does not provide clear instructions, is unfamiliar with the test materials or procedures for administration, and does not encourage the test taker, then the results will not be an accurate reflection of the learner's level of ability or skill. That it is why it is important for the test user to establish *rappport* with the test taker, to help the test taker to relax and to ensure that he/she is clear about the purpose of the test administration. Success on the practice items on a test can also be a good way of helping to reduce test anxiety and boosting confidence and motivation, together with encouragement from the test user to try hard and avoid careless mistakes. It goes without saying that the test user should be sufficiently practised in the administration of the test so that he/she can be attentive to the test taker's performance, and also that failure to adhere to the standardized instructions and scoring procedures will render the test results invalid.

It is also worth noting that a test taker's experience of tests can also have an effect upon his/her performance. In general, the more experience the test taker has of a particular test, the better their performance is likely to be, because of the effects of practice, reduced uncertainty (and hence anxiety) and increased confidence. This is why many publishers have guidelines for minimum intervals between re-administrations of tests. It can be useful to ask a test taker about their experience with a particular test before administering it.

Factors such as *gender*, *ethnicity*, *social class* and *disability* can also have an impact upon test results over and above the effects of the appropriacy of the format of the test, the medium of the test and the language used by the test which we considered in the last section. It is possible that test items may have a specific bias to the advantage or disadvantage of demographic groups. As Urbina (2004, p. 249) puts it:

An item is considered to be biased only if individuals from different groups who have the same standing on a trait differ in the probability of responding to the item in a specified manner. In tests of ability, for instance, bias

may be inferred when persons who possess equal levels of ability, but belong to different demographic groups, have different probabilities of success in an item.

This is known to test developers as *Differential Item Functioning* (DIF) and can result in systematic differences in average scores across different gender, social or ethnic groups, or between those with disabilities and those without, raising serious legal and ethical issues. This can arise from differences in the test takers' experiences of the content, as a result of *acculturation*, learning opportunities and experiences in both formal and informal settings.

It is important, therefore, to ensure that the developers of any instrument that is being considered for use have made the items in a test as fair as possible, without any content which may be offensive, inappropriate or even advantageous to a particular demographic group. Test materials should represent people from different ethnic, gender and social groups, disabled and non-disabled alike, in non-stereotypic as well as traditional roles and situations. Items that generate pronounced differences between demographic groups should be revised or deleted by the test developers. In this way, tests can take into account different acculturation and diversity in experiences to minimize any consistent underlying biases.

Identifying Sources of Information which can Corroborate the Information from a Test

The use of *multiple sources of information* in the interpretation of results to assist the decision-making process has already been highlighted as a characteristic of good practice in the use of tests. It is important that test results should be cross-validated with other evidence as a check on their validity, particularly if they will be used to support important placement or resource allocation decisions. The test manual or a review of the test will provide useful information in this regard and should be consulted by the test user. Test data may also be corroborated by means of:

- *existing school records*, which may include the results of other recently administered standardized tests or class-based assessments, or SATs;
- *examples of pupils' current coursework* including portfolios of classwork;
- *direct observation* of the learner's performance in class or other relevant setting (for example, the playground, as in the case of behaviour difficulties, or concerns about social relationships or bullying);

- *teachers' reports* on the learner's performance or behaviour; and
- reports from *assessments carried out by other professionals, such as speech and language therapists, educational psychologists, clinical psychologists and career guidance teachers* available as part of multi-disciplinary or inter-agency work.

Distinguishing between Different Types of Test: Tests of Attainment, Tests of Ability and Tests of Aptitude

The CCET (Level A) distinguishes between tests of *attainment*, tests of *ability* and tests of *aptitude*. Tests of attainment (or achievement) are the most frequently used tests in education. They have the following attributes:

- are designed to evaluate knowledge and understanding in curricular areas (for example, reading, spelling, number);
- may be individual or group-administered;
- may be used for screening or to determine progress or to evaluate the effectiveness of the curriculum;
- may include diagnostic tests;
- are generally norm-referenced, but may be criterion-referenced or curriculum-based; and
- are not designed primarily to be predictive.

Examples of tests of attainment may be found in Appendix 5 with details of publishers, and include such instruments as the Neale Analysis of Reading Ability – Second Revised British Edition (NARA II), the Wide Range Achievement Test – Expanded Edition (WRAT-Expanded), and the Numeracy Progress Tests Numeracy Baseline/Tests 1–6.

Tests of ability, on the other hand, are:

- designed to measure underlying general cognitive ability ('intelligence') or specific cognitive ability (for example, short-term memory, or verbal ability, or numerical ability) using tests of general knowledge, reasoning, problem-solving and memory rather than measures of what a test taker has learnt from his/her curricular experience; and
- in many cases, are designed to be predictive, to provide an indication of a learner's future academic performance.

Examples again are shown in Appendix 5, and include the Cognitive Abilities Test: Third Edition (CAT3), the Working Memory Test

Battery for Children (WMTB-C), and the Nonverbal Abilities Tests 6–13.

Finally, tests of aptitude are not widely used in education but are:

- tests of aptitude for practical tasks, such as musical aptitude, mechanical aptitude and spatial relations; and
- designed to be predictive.

Examples of tests of aptitude include the Primary Measures of Music Audiation (K–Grade 3) (Gordon, 1979) and Intermediate Measures of Music Audiation (Grade 1–6) (Gordon, 1982), both published by GIA Publications: Chicago.

Distinguishing between Different Types of Test: Formative and Summative Uses of Tests

A further distinction is made between *formative* and *summative* uses of tests. As William (2000) notes, formative test use is prospective in character, and may be regarded as *assessment for learning*. It is designed to provide:

- a source of effective feedback to the test taker about a gap between present performance and a desired longer-term goal;
- a means of actively involving the test taker in his/her own learning by gaining an understanding of his/her strengths and weaknesses; and
- information from the results of the test to help those teaching or supporting the test taker to identify ways of making teaching and learning more effective.

In contrast, summative test use is retrospective in character, and may be regarded as *assessment of learning*. It is designed to provide:

- a summary overview of what the individual knows, understands or can do at the time of testing; and
- a means of guiding formative test use.

Formative and summative test use are not different kinds of assessment but, rather, are different ways to which information arising from the assessment can be put. The outcome of the same assessment might serve more than one function and thus a given test can serve either a formative or a summative use depending upon the purpose of the assessment or the

assessment question. As William (2000) further points out, a key distinction is the extent to which test use utilizes information about *process* to provide suggestions for future action to help close the gap between present and desired levels of performance, or focuses upon the *outcome*. Thus a test of attainment in spelling could be used in a summative way if the purpose of assessment is to provide a standard score to summarize the test taker's performance relative to the norm group. But the same information could be used in a formative way if: (i) the data are used diagnostically to identify problems with particular spelling 'rules'; (ii) feedback is provided to the test taker regarding these difficulties; and (iii) teaching is modified to help him/her to learn these rules. See Black and William (1998, pp. 3–5) for a review of the evidence for the effectiveness of formative assessment use.

Distinguishing between Different Types of Test: Norm-Referenced, Criterion-Referenced and Curriculum-Based Measures

We considered the relationship between norm-referenced, criterion-referenced and curriculum-based measures and assessment needs in an earlier section. The CCET (Level A) requires test users to be able to distinguish between these three different types of measures:

- *Norm-referenced* measures interpret the learner's performance relative to that of the other individuals in a specific population, the normative sample of the test on whom the test norms are based in terms of *developmental scores* (for example, age-equivalent scores, such as 'Reading Ages' or 'Spelling Ages') or *derived scores* (for example, standard scores, or percentiles).
- *Criterion-referenced* measures interpret the learner's score in terms of an empirical, statistical relationship between the test score and an external specific, objective, pre-defined standard, or *criterion* (for example, pass/fail, right/wrong, number correct, percentage correct, or number of correct responses per minute).
- *Curriculum-based* measures interpret test scores in terms of specific aspects of the content of the learner's existing curriculum usually in terms of specific instructional objectives.

By way of caveat, it should be noted that the quality of norm-referenced tests is dependent upon the adequacy and representativeness of their standardization, and on their reliability and validity (which we will discuss in more detail in Unit 2). The quality of criterion-referenced and, by extension,

curriculum-based tests, is dependent upon what constitutes an acceptable level of performance for the attribute being measured. Should the criterion be 100 per cent or 80 per cent correct, for example? And should this criterion change as a function of the test taker’s cognitive abilities, gender, social class or ethnicity?

Completing the Unit 1 Competency Assessments

We shall now see how this fits in with the BPS competences and the answers required by the assessment that you will be required to undertake. Only short-term assessments are required for Unit 1, but two pieces of evidence are required for each element. The examples provided utilize short note answers, together with opportunities for self-assessment using multiple choice questions, and simulation exercises from a test publisher’s catalogue. But your assessor may provide you with case studies or scenarios and ask you to provide a commentary. Note that ‘1.1’ below refers to the assessment for Unit 1 Element 1 and ‘1.2’ to Unit 1 Element 2 and so on. The competences are shown in full in Appendix 1, and the *Guidance for Assessors* material is reproduced in full in Appendix 2.

Examples of How the BPS Competences might be Assessed and Answered

1.1 Which of the following assessment needs would you consider a test to be ‘highly appropriate’ in addressing?

Assessment need	Indicate ‘Yes’ or ‘No’
To compare the learner’s knowledge base with the norm for his/her age	Yes/No
To assess the content of a learner’s knowledge base within a specific domain	Yes/No
To generate hypotheses about how the test taker learns	Yes/No
To investigate how responsive the learner is to attempts to intervene	Yes/No
To assess what may be interfering with the learner’s ability to benefit from existing levels of instruction	Yes/No
To determine the amount of time a learner spends ‘off-task’ in class	Yes/No

(Check your response with the answer shown in Appendix 6.)

1.1 What is the relationship between assessment and testing?

Assessment is a process of collecting data for the purpose of making decisions about individuals and groups by means of:

- *Direct observation* of the student/pupil/learner's performance or behaviour.
- *Existing information*, such as portfolios of the learner's work.
- *Interviews* with those who know the learner well, such as former teachers or parents also provide useful information.
- *Professional judgements* of those who have worked with the learner in the past or are currently working with the learner in the here and now.
- A *test* may be used.

Testing is thus a sub-set of assessment.

1.2 Which assessment needs might best be addressed by the use of a test procedure?

In general, assessment needs relating to entitlement or eligibility for additional support or accountability might best be addressed by means of a test. Examples would include:

- How effective is the cued spelling programme in the KS2 class?
- Should a student receive special arrangements in statutory examinations?
- How many pupils in the Year 1 intake have poor levels of phonological awareness?
- Does a pupil require additional learning support to address literacy problems?

In contrast, assessment needs relating to instruction or where use of a test would be problematic because of an inappropriate norm group, disability, or cultural bias, might be more appropriately addressed by means of an alternative assessment approach, such as observation, criterion-referenced measures, curriculum-based assessment or dynamic assessment. Examples would include:

- Has a pupil mastered addition and subtraction number 'facts' to 20?
- How can a pupil most effectively learn to spell correctly words misspelled in an essay?
- Can a pupil read a given text with 90 per cent accuracy?

- Does a pupil for whom English is an additional language and who is newly arrived in the UK have a problem in receptive vocabulary?

1.2 Complete the following table to match assessment approach to assessment need

Indicate which assessment needs would best be addressed by the use of a test procedure and those which might be more appropriately addressed by means of an alternative approach.

Assessment approach	Assessment need ^a		
	Entitlement/eligibility	Accountability	Instruction
Norm-referenced			
Criterion-referenced			
Curriculum-based			
Dynamic assessment			

^a ‘√/√’ denotes highly appropriate, ‘√’ appropriate and ‘X’ not particularly appropriate

(Check your response with the answer provided in Appendix 6.)

1.3 Select a specimen test or a test from a catalogue.
Use your knowledge of: (a) child development; and
(b) the skills being assessed, to support its use

The Neale Analysis of Reading Ability – Second Revised British Edition (NARA II) is an individually administered, norm-referenced timed test of oral reading. It consists of two parallel forms designed to assess the rate, accuracy and comprehension of reading of children aged between 6 and 12 years, together with supplementary diagnostic tests of phonemic awareness, memory for regular and irregular word patterns and auditory discrimination.

The manual provides detailed instructions for the administration and scoring of the test together with background information about its development. There is also a demonstration cassette with examples of prompting and of recording errors and a commentary highlighting good practice and possible problems that the test user may encounter. The test was standardized in 1996, so the norms are less than 10 years old and are based upon 1,928 children for Form 1 and 1,546 for Form 2 from across the UK.

The norms are also weighted for gender and for social class, using free school meals.

The test consists of graded passages, or narratives, presented in book form. Each narrative is a self-contained story and is accompanied by a picture, which sets the scene but does not provide details of the story. Accuracy is assessed by recording the number of errors and rate of reading by calculating the number of words correctly read per minute. Comprehension is assessed by questions about the main idea of the narrative, the sequence of events, recall of details and questions that also require the test taker to draw inferences about the material read. Separate norms for accuracy of reading, rate of reading and comprehension are provided.

The level of difficulty and interest levels of the narratives and questions are appropriate for the 6–12 year age-range. The format of the test, instructions and materials are also appropriate for a test of oral reading and there are no problematic cultural references. Pupils respond orally by reading a passage aloud and then answering standard comprehension questions asked by the test administrator. This is appropriate for a test of oral reading. Provision is made in the instructions for helping test takers deal with words in the narratives that they may be unfamiliar with. The test user is allowed to correct errors or to supply words which the pupil does not know.

A practice passage is provided to ensure that the test takers know what is expected of them, and there are clear instructions for establishing a basal level, the level below which it is assumed that the test taker will get all of the items correct (here, no more than two errors), and a ceiling level, the point at which it is assumed that the test taker will fail all items and testing is discontinued (here, no more than 16 errors are made on passages 1–5 or no more than 20 errors on passage 6). Administration of the test should thus take no more than 20 minutes, which is appropriate for pupils in the target age-range. There is also an extension passage for use with older pupils or more able readers within the age-range, with more demanding vocabulary and concepts.

In conclusion, this is an attractively produced test of reading which is well-suited to the purpose of assessing the reading skills of pupils in the 6–12 year age-group.

1.4 Describe how measures of ability and attainment of students/pupils/learners are influenced by environmental factors and give specific examples

Factors such as temperature, poor lighting and noise can affect the test taker's concentration and ability to attend and hence his/her performance

on measures of ability and attainment. Overcrowding may also make it difficult for the test taker to hear instructions and can also provide sources of distraction, and an increased likelihood of copying answers or cheating.

The test user should ensure that the room used for testing is adequately ventilated, with an appropriate background temperature, and is well-lit, free from extraneous noises and from other possible distracters (such as food or drink, or third parties who might influence the test taker's response), with appropriate seating and adequate space for test takers (for example, a desk and chair of the correct height, and well-spaced-out desks or tables for administration of a group test to minimize copying of answers). The session should also be free from interruptions.

Tiredness and lack of motivation, for example, can adversely affect the test taker's performance, as well as test anxiety. Test anxiety is a fear of being evaluated, or fear of failure, or anxiety about the context in which testing takes place or perhaps even about the specific type of test used (for example, a timed test or perhaps a test of mathematics). If severe, test anxiety can have an adverse effect upon the test taker's performance on tests of ability or attainment.

Variables associated with the test administrator can also affect the results obtained. If the test user does not establish rapport with the test taker to relax him/her and ensure that they are clear about the purpose of the test session, does not administer the test properly following the standardized instructions, does not provide encouragement to the test taker, and does not score the results accurately and in accordance with the guidance in the test manual, then the results would be invalid.

Finally, factors such as *gender*, *ethnicity*, *social class* and *disability* can also have an impact upon measures of ability and attainment at two levels. First, they can have an effect at the level of the appropriacy of the norms for specific demographic groups defined in terms of gender, ethnicity, social class and disability, the appropriacy of the level of the items (both in terms of the developmental stage of the test taker and interest level), the accessibility of the test in terms of its format, the appropriacy of the modes of presentation and of responding, the language used by the test, and also the learner's exposure to the content being tested. Second, it is possible that test items may have a specific bias to the advantage or disadvantage of certain demographic groups.

To illustrate this, tests presented orally to test takers unable to understand the instructions and task demands (for example, those with severe hearing impairments, or with severe language and communication problems, or those for whom English is an additional language) would be unfair. A learner with a severe speech and language delay or disorder may

have difficulty in providing a clear verbal response, and a learner with a physical disability or motor difficulty, such as dyspraxia, may have problems responding in writing or via the manipulation of objects, particularly if the test has time limits. Similarly, tests that assume that the test takers are able to read items and instructions independently and to record their responses in writing could be unfair for those with marked visual impairment, and for those with literacy or motor skill problems. A learner with a poor record of school attendance, or perhaps a test taker recently relocated from a different country or culture would be at a disadvantage in regard to a test of a curricular area if he/she had missed the opportunities for instruction afforded to regular attenders at school. Test results in such cases might be held to reflect the effects of opportunities to learn rather than the skill or ability being tested.

In addition, the test materials should neither give offence (for example, text or pictures of pigs in the case of some cultural and religious groups) or an unfair advantage or disadvantage to test takers from a particular gender, religion, ethnic group, social class or disability in the form of systematic differences in average scores across such demographic groups. This can arise from differences in the test takers' experiences of the content, as a result of *acculturation*, their learning opportunities and experiences in both formal and informal settings. Test materials should represent people from different ethnic, gender and social groups, disabled and non-disabled alike, in non-stereotypic as well as traditional roles and situations. Items which generate pronounced differences between demographic groups should be revised or deleted so that tests can take into account different acculturation and diversity in experiences to minimize any consistent underlying biases.

1.5 Give examples of the kinds of information which may be used to corroborate the information elicited by a test or other assessment technique

Test data may be corroborated by means of:

- *existing school records*, which may include the results of other recently administered standardized tests or class-based assessments, or SATs;
- *examples of pupils' current coursework* including portfolios of classwork;
- *direct observation* of the learner's performance in class or other relevant setting (for example, the playground, as in the case of behaviour difficulties, or concerns about social relationships or bullying);
- *teachers' reports* on the learner's performance or behaviour; and

- reports from *assessments carried out by other professionals, such as speech and language therapists, educational psychologists, career guidance teachers* available as part of multi-disciplinary or inter-agency work.

1.6 Explain the differences between tests of attainment, tests of ability and tests of aptitude

Tests of attainment (or achievement) are the most frequently used tests in education and:

- are designed to evaluate knowledge and understanding in curricular areas (for example, reading, spelling, number);
- may be individual or group-administered;
- may be used for screening or to determine progress or to evaluate the effectiveness of the curriculum;
- may include diagnostic tests;
- are generally norm-referenced, but may be criterion-referenced or curriculum-based; and
- are not designed primarily to be predictive.

Examples of tests of attainment include such instruments as the Neale Analysis of Reading Ability – Second Revised British Edition (NARA II), the Wide Range Achievement Test – Expanded Edition (WRAT-Expanded), and the Numeracy Progress Tests Numeracy Baseline/Tests 1–6.

Tests of ability, on the other hand, are:

- designed to measure underlying general cognitive ability ('intelligence') or specific cognitive ability (for example, short-term memory, or verbal ability, or numerical/quantitative ability) using tests of general knowledge, reasoning, problem-solving and memory rather than measures of what a test taker has learnt from his/her curricular experience; and are
- in many cases, designed to be predictive, to provide an indication of a learner's future academic performance.

Examples of tests of ability include the Cognitive Abilities Test: Third Edition (CAT3), the Working Memory Test Battery for Children (WMTB-C), and the Nonverbal Abilities Tests 6–13.

Finally, tests of aptitude are not widely used in education but are:

- tests of aptitude for practical tasks, such as musical aptitude, mechanical aptitude and spatial relations; and
- designed to be predictive.

Examples include the Primary Measures of Music Audiation (K–Grade 3) (Gordon, 1979) and Intermediate Measures of Music Audiation (Grade 1–6) (Gordon, 1982), both published by GIA Publications: Chicago.

1.6 Consult the publishers’ catalogues and indicate which of the following are tests of attainment, which tests of ability and which tests of aptitude

Test:	Test of attainment	Test of ability	Test of aptitude
Naglieri Nonverbal Ability Test – Individual Administration (NNAT) (http://www.harcourt-uk.com)			
Wechsler Individual Achievement Test – Second UK Edition (WIAT-II) (http://www.harcourt-uk.com)			
SPAR (Spelling & Reading) Tests A & B (http://www.hoddertests.co.uk)			
Verbal Abilities Tests 6–10 (http://www.hoddertests.co.uk)			
Music Aptitude Profile (Grades 5–12) (http://www.giamusic.com)			
Raven’s Progressive Matrices & Vocabulary Scales (http://www.nfer-nelson.co.uk)			
Primary Reading Test (Levels 1 & 2) (http://www.nfer-nelson.co.uk)			
Modern Language Aptitude Test – Elementary (http://www.2lti.com/mlate.htm)			

(Check your response with the answer provided in Appendix 6.)

1.7 Explain the difference between formative and summative uses of tests, giving examples of each type

Formative test use is prospective in character, and may be regarded as *assessment for learning*. It is designed to provide:

- a source of effective feedback to the test taker about a gap between present performance and a desired longer-term goal;

- a means of actively involving the test taker in his/her own learning by gaining an understanding of his/her strengths and weaknesses; and
- information from the results of the test to help those teaching or supporting the test taker to identify ways of making teaching and learning more effective.

Summative test use is retrospective in character, and may be regarded as *assessment of learning*. It is designed to provide:

- a summary overview of what the individual knows, understands or can do at the time of testing; and
- a means of guiding formative test use.

Formative and summative test are not different kinds of assessment, but rather different ways to which information arising from the assessment can be put. The outcome of the same assessment might serve more than one function and thus a given test can serve either a formative or a summative use depending upon the purpose of the assessment or the assessment question.

A test of attainment in reading or spelling would be used in a summative way if the purpose of assessment was to provide a standard score to summarize the test taker's performance relative to the norm group.

A test would be used in a formative way if the data were used diagnostically to identify problems, feedback provided to the test taker regarding these difficulties, and teaching modified to help him/her to learn these rules.

1.8 Differentiate between norm-referenced, criterion-referenced and curriculum-based measures

- *Norm-referenced* measures interpret the learner's performance relative to that of the other individuals in a specific population, that is, the normative sample of the test on whom the test norms are based, in terms of *developmental scores* (for example, age-equivalent scores, such as 'Reading Ages' or 'Spelling Ages') or *derived scores* (for example, standard scores, or percentiles).
- *Criterion-referenced* measures interpret the learner's score in terms of an empirical, statistical relationship between the test score and an external specific, objective, pre-defined standard, or *criterion* (for example, pass/fail, right/wrong, number correct, percentage correct,

or number of correct responses per minute). A criterion-referenced measure is used to measure mastery of information or skills, what the learner can or can't do in terms of absolute standards.

- *Curriculum-based* measures interpret test scores in terms of specific aspects of the content of the learner's existing curriculum usually in terms of specific instructional objectives.