

Assessment Policy

1 Introduction

This document has been written by the National Foundation for Educational Research (NFER) to provide policy makers, researchers, teacher educators and practitioners with an overview of the NFER philosophy towards assessment. This philosophy underpins our approach to assessment development, assessment research and developing policy papers. The Policy describes NFER's wealth of experience in the assessment environment in the UK, the principles that underpin our work in this area, and some of the processes that are used in our assessment development projects. The term assessment is used to encompass the full range of approaches, from formative to summative and for use in formal and informal contexts.

2 Background

Since 1946 NFER has been working to provide the most innovative thinking, practical research and responsive assessment programmes to underpin the imperatives of the time. NFER has worked closely with government bodies (currently these include the Department for Children Schools and Families, the Qualifications and Curriculum Authority and Ofqual in England, the Scottish Government, and the Department for Children, Education, Lifelong Learning and Skills in Wales), local government agencies (in particular the Local Government Association and local authorities), and universities. NFER's work enables policy makers and practitioners to make better, more informed decisions, drawing on sound evidence and accurate information.

Throughout its history NFER has been involved in a wide range of assessment initiatives: initially playing a large role in developing and analysing Eleven Plus selection tests, being actively involved in national monitoring systems from the first national surveys of reading in the 1950s, then with the Assessment of Performance Unit from 1974 to 1988, and in the international surveys (TIMSS, PIRLS and PISA¹) up until the present time. NFER was a lead organisation in developing and piloting the first National Curriculum assessments in England from 1989 and has continued to develop many of the tests since that time, including English, maths and science tests. NFER has recently set up a subsidiary company, *i-nfer*, to research and develop assessments making the best use of new technologies.

With regard to examinations for older learners and adult learners, NFER has conducted studies measuring the usefulness of A Level in predicting achievement on university courses and is currently involved in an evaluation of an aptitude test for similar purposes. NFER has also been involved in the formal evaluations of the Diploma programme and developing materials to support Skills for Life assessments.

¹ The Trends in International Maths and Science Study, the Progress in International Reading Literacy Study, and the Programme for International Student Assessment.

This list of projects is included to exemplify the range of NFER's experience and expertise, and to demonstrate the central role which NFER has played in the testing and examinations system in the United Kingdom for over 60 years.

NFER has achieved this position by always aiming to conduct high quality research, to develop high quality assessment instruments and by drawing on our vast knowledge and experience of the curriculum and its assessment to inform developments wherever possible. As an independent research organisation the Foundation is able to use its position to provide advice and shape projects without influence from political agendas or other initiatives.

The NFER view of assessment is to acknowledge and embrace the variety of assessment purposes and processes that are available, and we believe many models have their place in the overall assessment enterprise. Our work embraces both formal and informal assessment, and internal (teacher assessment) and external assessment, and it is important to recognise the distinctive features and requirements of each. The best approach to this is to understand and accept the distinctions and relationships between the different forms of assessment, and to give appropriate attention to each one.

Our stance in relation to assessment is that there must be a clear statement of the intended purpose(s) of the assessment system and that its processes and instruments should have an appropriate amount of evidence of validity and reliability to provide sound support for those intended purpose(s). This implies that there should always be a thorough development cycle for instruments, and evaluative research to demonstrate that soundly-based judgements are being reached on the basis of the assessment system. The rest of this policy document goes on to give more detail about these different aspects.

3 Assessment Development Fundamentals

To ensure the high quality of an assessment a number of factors need to be put in place before any development work starts. First, the purpose(s) to which the results are going to be put need to be established. These purposes will determine the most appropriate model for the assessment, the nature of the development process, the style and content of the assessment, the analysis required and so on. Purposes vary greatly, but they can include:

- collecting evidence of student learning on a particular topic so that immediate feedback can be given which the learner can use to progress to the next level;
- measuring the performance of a given population of learners in relation to a given curriculum (a summative assessment of what the learners have or have not achieved over a course or period of time);
- monitoring different levels of achievement over a period of time;
- measuring students' potential for future learning or achievement (such as cognitive aptitude tests).

Once the purposes are agreed it is essential to define the construct that will be measured, that is the skills, knowledge and understanding that the learners are expected to have developed in

order to be successful in the assessment. In most cases the questions or tasks in the assessment will assess a sample of the construct, including content and processes, and use the results to generalise achievement across the whole area. The construct must be carefully designed to link to the agreed purposes. This linking is essential for the validity of the final assessment and the uses of its results. (Validity is discussed in more detail in the section on The Assessment Development Process.)

Another fundamental is the definition of the target group to be assessed. The level of skill and knowledge of the cohort will determine the level of difficulty of the assessment. Similarly, the range of their ability will determine the design of the assessment in terms of the balance across questions of different levels of difficulty. In other words, the target audience for the assessment should be closely linked to the design of the assessment itself.

Once these factors are agreed it is possible to define the test or assessment specification. This should detail the number and nature of questions to be included, as well as the balance across content knowledge and skills, and the difficulty levels of the questions. In a less formal assessment, the nature of the tasks and of the assessment judgements should be specified.

To ensure that a high quality assessment is developed, an expert team must be put in place. This should include subject matter experts, assessment experts who can advise on constructing the assessments, and (where applicable) statisticians who can analyse the results. The team should also include experts who can advise on the accessibility of the assessments for all potential groups of users, and advise on any possible biases against particular groups.

Finally, to make sure that assessments are used in the way that they were designed, and that the information is used to the best advantage of the students, teachers trained in the effective use of assessment and its results are needed.

4 Dimensions of Assessment

In broad terms, educational assessment serves two major purposes. Firstly, to provide immediate feedback to teachers and learners, in order to facilitate the learning process. This is often termed formative assessment, but may also be referred to as diagnostic assessment or, more recently, as Assessment for Learning. The second major purpose is to provide information which summarises a particular phase of learning or education. This information may be for a variety of institutional uses such as monitoring progress, certification or selection, and it may be for a variety of users, such as parents, teachers, governmental bodies and the learners themselves. This type of purpose is often termed summative assessment. Both these purposes are important in the educational system as a whole, but they have different requirements and different characteristics. However, the fundamentals described above apply to both types of assessment.

This dimension of purpose is only one categorisation. A different categorisation, which cuts across this, describes formal and informal processes of assessment. The distinction here is between, on the one hand, formal processes such as exams, tests and other assessments in which learners encounter the same tasks in controlled and regulated conditions and, on the other hand, those less formal activities that form part of on-going teaching and learning. This

second group would encompass question and answer, teacher observations, group discussion and practical activities together with classroom and homework writing and assignments.

Using this two-fold classification (as shown in the table below), it can be seen that formal and informal assessments can each be used for both formative and summative purposes. The four cells of the table depict the key aspects of assessment systems and instruments.

Processes	Purposes	
	Formative	Summative
Informal	<i>Questioning</i> <i>Feedback</i> <i>Peer assessment</i> <i>Self assessment</i>	<i>Essays in uncontrolled conditions</i> <i>Portfolios</i> <i>Coursework</i> <i>Teacher assessment</i>
Formal	<i>Use of test results</i> <i>Further analysis of tests, exams, essays</i> <i>Target setting</i>	<i>Tests</i> <i>Exams</i> <i>Essays in controlled conditions</i>

Informal Formative Assessment

Formative assessment is vital to teaching and learning. It is embedded in effective classroom practice and it should be tailored to the individual and his or her own stage of learning. Such processes have long been regarded as essential for progress, providing a motivational effect for learners as well as information on what has been recently achieved and the next steps to be taken in order to make progress.

Although such practices have always been intrinsic to successful teaching, recent research and policy has characterised them as a particular approach to assessment, leading to the principles that have been set out under the heading of ‘Assessment for Learning’. Its characteristics include: sharing learning goals with the learners, helping learners know and recognise the standards they must aim for, providing feedback that helps learners to know how to improve and the inclusion of both self and peer assessment.

NFER are very supportive of the principles of Assessment for Learning and believe these must be further promoted and new and better resources must be produced and supplied to teachers. There is also a need for further, and more rigorous, research which explores the successful elements of Assessment for Learning in practice.

Formal Formative Assessment

Formal assessments can also be used for formative purposes, whenever an analysis is made of performance in a formal assessment and the results used to plan teaching for classes, groups or individuals. As an example, the end of key stage tests in England over the last few years



have been systematically analysed in order to draw out implications for teaching and learning.

A major focus of NFER's current work is the formative use of assessment information gained by formal means. Since 2005, we have been researching the potential of e-assessment in low-stakes contexts to support Assessment for Learning, and a formative e-assessment product known as '*i-nfer* plan' is now available for primary schools.

Informal Summative Assessment

Since teachers have, by the very process of teaching, a wealth of informal assessment information on each learner, there is a strong incentive to find ways of summarising that information so that it serves a summative purpose. Ongoing informal assessment outcomes cover learners' performance across a range of contexts, and are thus potentially both more valid and more reliable than a result from a single formal assessment.

Our view is that results from informal assessments can be used to provide summative information. However, there are three conditions that must be fulfilled before it can be introduced successfully. Firstly, a major investment has to be made in teacher professional development in order to bring about a shared understanding of criteria. Secondly, a part of this professional development would need to address teachers' and advisers' understanding of the nature and purposes of the four quadrants of assessment, as shown in the table above. Finally, the system would need an element of external monitoring and accountability that commanded public and professional confidence.

Formal Summative Assessment

Formal summative assessment can serve many purposes. Among these are reporting to parents (from school-based end of year tests), school accountability (as with end of key stage 2 tests in England), certification of schooling (as with GCSE) and selection (as with A Levels for university entrance). The purposes range from providing individual information to grouped information. They also provide information on personal accountability and system accountability.

NFER has been involved in the development of national curriculum tests since their introduction, and these tests provide a useful example of the range of purposes that formal test results are put to. These national curriculum test results are used for many, sometimes contradictory, purposes, and NFER believe that the current tests serve the required purposes as well as any single measure could. However, in an ideal world no one test would be required to meet so many purposes.

5 The Assessment Development Process

Many different processes are in use in the UK and in international settings for assessment development. The nature of the assessment development process will vary according to many factors, including:

- the importance placed on the final results; if the assessments are high stakes then the processes will be more complex and are likely to demand high security;
- the type of questions or tasks within the assessment, so different processes are appropriate for essay questions as compared to multiple choice questions;
- the assessment culture within a country, in that more faith may be given to professional judgement during the process, compared to statistics, or there may be more or less teacher involvement in the process.

A key aspect of any assessment is that there is good evidence that its results can be used in a valid way, that is, they can be used for the purpose(s) for which they were designed. A critical aspect of validity evidence is that the test itself is reliable. One key aspect of reliability is that the same result would be achieved by a learner if the assessment were administered on a different occasion, or if a different version of the assessment were used. A second key aspect of reliability is the extent to which the questions within the assessment all measure the same feature; this is usually termed internal reliability. For there to be evidence that assessment results are valid there must be accurate information about what the assessment is designed to measure and the results must be used for the purpose(s) for which they were designed. At each stage of the assessment development process the extent to which the assessment is valid and reliable should be gauged. An assessment that has good evidence of validity is fit for its purpose.

The time necessary for the assessment development process will vary, again depending on the nature of the assessment. National Curriculum tests in England, for example, as developed by NFER, generally take three years to develop, although some assessments take much less time than this. The timing is dependent on the number of stages in the process, as well as the way in which the process is designed – for example National Curriculum test development processes involve a second pre-test one year in advance of the live test usage, so that learners of the right age, and at the right stage of their learning, can be involved.

Question writing will generally be the first stage of any assessment development (in the development of reading tests it will be preceded by text selection). Question writing must be guided by a clear assessment specification, determined by the assessment's nature and purpose. It is important that writers are expert in the subject being assessed, at the right level, and trained in assessment development. A number of key rules should be followed when writing questions, such as the use of simple sentences and active verbs wherever possible. NFER works to tailored guidelines developed over a number of years.

In many cases after question writing, there will be a process of informal **trailing**. This involves trying out the questions with a small group of learners to ensure that they are assessing what they are intended to assess in the way that was intended. This can be part of the process for all types of questions and for assessments for many different purposes. In some cases this may adopt a 'cognitive laboratories' approach in which learners 'think aloud' whilst answering the test questions enabling the developers to review what is really being assessed.

After informal trialling, usually the questions will be combined into test booklets or, in the case of computer-based assessment, programmed into a test delivery engine. They are then submitted to one or more formal trials, also known as field tests, field trials or **pre-tests**. Pre-tests can provide a great deal of information to inform the development process, including how difficult the questions are, whether the questions discriminate between learners of different ability, whether the questions work together as a whole assessment, and to inform the development of the mark schemes. Pre-tests also provide information for setting pass marks or grade thresholds from the assessments. For longer questions with a higher mark allocation the purpose may be confined to ensuring that the questions are functioning appropriately and informing the mark scheme development.

NFER works closely with schools and other centres in the development of assessments and values their input highly. In all assessment development processes, **expert review** is a key part of the process. The professional judgement of content experts is essential for ensuring that the questions are set at an appropriate level, assess the full range of the curriculum, and assess a range of different skills. The requirements against which the experts are making their judgements will depend on the assessment specification.

Informal assessments, for example classroom activities with assessment checklists, do not require large-scale pre-tests or statistical analyses. However, school trials are nevertheless important to check their appropriateness and manageability.

An integral part of the development process for new assessments should be a **pilot**. This should involve using the assessments in a live situation with the target group of learners to ensure that they work as expected, and that there are no immediate unintended consequences (it is unlikely that the pilot will run for long enough to ensure that all unintended consequences are picked up).

Finally, all assessment systems should have an **evaluation** built into the delivery process. It is essential, especially for new tests, that they are evaluated against the agreed purposes. As with the test development processes the evaluation processes should be fit for purpose. More detail is given on this in the section on evaluation below.

6 Standard Setting

As with assessment development, the standardisation of assessments takes different forms depending on the nature of the assessment outcomes and purpose(s). Perhaps the most straightforward form of standardisation is called norm referencing, in which the levels are set as a result of the percentage of learners achieving a certain mark, for example the top 5% being awarded the highest grade. Other forms of standardisation include by age, so that age standardised scores are produced stating what level(s) of achievement is expected of learners at a certain age. More sophisticated techniques can be used in e-assessments, in which complex statistical procedures can be used to categorise performance of different types, and link this to formative feedback. In this section we will focus on the techniques used for making grade or level judgements in formal tests.

Standard setting and maintaining standards over time in national curriculum tests or external examinations, such as GCSE or A Level, are perhaps the most demanding and complex parts of assessment development. This is also an area where there is much public debate and media interest. A number of possible methods can be used for standard setting, and NFER uses all of these to a greater or lesser extent. It should be noted that no one way is ‘the right way’, and that different methods may lead to slightly different results.

The NFER approach is to use a range of methods and to consider the information that each method provides. A balanced and pragmatic view is taken of the different sources of evidence to come up with a final answer. Methods for setting standards initially include:

- professional judgement: to consider what it is that the assessment is aiming to achieve, what level of success is appropriate for an assessment of this nature, and how many of the questions in the assessment a learner at different levels of ability should be expected to succeed at. It is likely that a review of learner scripts will be a key part of this process;
- statistical analysis: statistical information should be available on the performance of different questions in the test, and for the test overall. This will provide information to support the professional judgements; for example whether learners at different levels of ability answer certain questions correctly or not. Statistical analysis can also be used when an external measure is available against which the standard can be set or when parallel forms of an assessment, with common questions, are used. It can then be used to scale performance or to link performance on a new test to performance in an earlier test, or to an anchor test at an agreed standard.

The methods for maintaining standards over time are likely to be similar to those for setting standards. In this case the information will be used to compare the performance of the current test to parallel versions of the test from the previous years.

- Professional judgement including script scrutiny is likely to form part of the process where scripts from one year, with a known level, are compared to scripts from the current session, so that equivalent levels of achievement can be agreed.
- Statistical analysis is used in this context to produce information for equating marks on one test with marks on another test, as well as to verify that the test has performed as expected, and in a similar way to previous versions of the test. Measures of internal reliability of the test will be useful.

As described above it is likely that none of these sources of information will provide one right answer, and NFER believes it is important that information from various sources is collected, and reviewed, to ensure an informed decision is made.

7 Results

Results from assessments can be reported in many different ways linked to the purposes and the nature of the assessments. As with the assessments themselves, results can be more or less formal, taking the form of a conversation between a learner and his or her teacher, or a

certificate summarising achievement to date that can be used for selection purposes. Results can also be for many different audiences: individual results for the learner, class results for the teacher or school leaders, whole year group results for school leaders, local authorities or national monitoring, or even a collation of results by local authority or for a whole cohort.

The results from assessments also vary greatly in the level of detail provided. For end of key stage tests, results are reported by level; for GCSEs or A Level they are reported by grade. This level of reporting is fairly broad, and for a high stakes test, this is necessary as it is possible to be reasonably confident about the reliability of the results for individual learners. This degree of certainty is essential for high stakes tests where the results are to be used for selection purposes or for school accountability.

For other forms of assessment it is possible to provide more detailed feedback or results than just an overall level or grade. Where the results do not require such a high degree of reliability for individual learners it is possible to provide feedback at the sub-test level, for example by curriculum area. As each of the sub-tests will have fewer questions than the full test each of the judgements will be less reliable than the overall result but may still be high enough for a low stakes assessment environment. Where results for a group of learners or a group of questions can be collated it may again be possible to report in more detail, as the larger number of results will increase the overall reliability of the judgements. For example, results from the national curriculum optional tests in England are used to provide detail of strengths and weaknesses in different curriculum areas for individual learners or groups of learners.

The use of technology in assessment makes it much easier to provide more detailed information to learners and teachers, again especially where this is to be used in a low stakes environment. In the *i-nfer* e-assessments, the assessments have been calibrated in such a way that typical profiles of performance have been defined. As learners take the tests their performance is mapped against these profiles and implications for future teaching and learning are detailed.

In England there is currently a debate about publishing information about the extent of error in the results of a given assessment alongside the results. For example, it would be possible to report that the results in a 100-mark test are accurate to plus or minus 3 marks. This would mean that those using the results would have more information on which to base their judgements. Error levels are already published alongside value added measures in England, and much more widely for test scores themselves in the United States. NFER believes that the provision of such information is useful for the users of assessment results.

It is frequently the case in education systems that information about performance collected for one purpose is then thought to be useful for additional purposes, for which the assessment was not originally designed. In the NFER submission to the Education Select Committee on Testing and Assessment in England seven different purposes for the end of key stage tests were given. The assessments and the results of these assessments were not originally designed to serve all these purposes, some of which have emerged over time, calling into question the validity of decisions made against these purposes.

As with all other aspects of the assessment process, when considering the nature of results and reporting it is essential to consider the original purpose(s) of the assessment. Where results are to be used as part of a selection process, it may or may not be beneficial to have information about strengths and weaknesses on different curriculum areas, but information about groups of learners is likely to be less useful than the individual results. In such a context it is likely that the accuracy of the measure, that is the error of the measurement, will provide useful additional information. Where an assessment is used part way through a programme of learning, then it is likely that information about strengths and weaknesses of individuals and groups of learners will be useful in order to inform planning of future lessons.

8 Monitoring and Evaluation

When any new assessment programme is introduced, a process for evaluating the assessment should be introduced at the same time. A thorough evaluation is likely to include: statistical analysis to ensure that the assessments are functioning in a reliable way and the results can be used to serve the purposes for which they are required; and qualitative means of collecting broad as well as in-depth views of the various stakeholder groups. There should be the opportunity for the findings from the evaluation to feed back into any further development of the assessment.

Throughout this policy document the importance of agreeing a clear purpose for any assessment has been emphasised. As described above an important part of the overall process is an evaluation of whether the final assessment, when used in the live context, provides results that can be used for the agreed purpose. It is also important to monitor whether the results are being used for purposes for which they were not intended.

Unintended consequences of assessment can have a huge impact on learners and on teachers and no matter how much care is taken in the development and piloting of a new assessment, it is only when it is introduced and used in a live context that many of the consequences become apparent. An evaluation is a valuable means of picking up on these consequences.

If at all possible an evaluation should be a long-term endeavour, in which impacts of the assessment on the system (for example changes to the curriculum or changes in approaches to teaching and learning), changes to standards, and uses and purposes of the results can all be monitored over time, and in particular once the assessment becomes established and embedded in the system.