

assessment literacy

for wise decisions

A publication commissioned by the Association of Teachers and Lecturers
from Sue Swaffield, University of Cambridge Faculty of Education
Pete Dudley, National College for School Leadership Networked Learning Group



FOREWORD

It is a sign of the times that barely a year after *Assessment literacy for wise decisions* was produced, it has proved to be one of ATL's most popular publications. The heightened emphasis on testing and assessment, and the production of increasing amounts of complex data mean that, more than ever before, practitioners need to get to grips with the possibilities and limitations of assessment information.

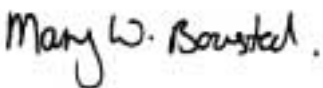
The feedback we have received from those who have used this guide confirms that practitioners and schools have found it extremely helpful in understanding assessment processes and the language of assessment. Schools using the guide as the basis for INSET tell us that it really has helped them to develop 'assessment literacy' – an understanding of the issues surrounding assessment and performance data.

We owe a great debt to the authors, Sue Swaffield, a lecturer at Cambridge University's Faculty of Education, and Pete Dudley, now working on networked learning communities for the National College for School Leadership. They carried out the work with great intelligence and enthusiasm, and provided a huge amount of wisdom and first hand experience. They can feel justifiably proud that the Scottish Executive has approached ATL for permission to publish a co-edition of this guide for schools in Scotland.

This second edition is being published at a critical time. Six years on from publication of Labour's election manifesto in 1997, important questions are being asked about whether the Government's mission to raise standards, and its heavy-handed, top-down approach, is working in the best interests of pupils and schools. Through testing we communicate standards, but has the emphasis on the 'three Ts' of targets, tests and tables gone a step too far – and in the wrong direction?

There are already straws in the wind about a major policy shift away from external testing and towards more teacher assessment. There are also signs that ministers are beginning to ask deeper questions about assessment for learning, and the role of teacher assessment in learning. Of course a move towards the greater involvement of teachers in assessment and testing would beg a number of important questions, not least about teacher workload. But it is, I believe, a step in the right direction.

Whatever the political imperatives, it is always important for the educator to remember that assessment starts with learning, it supports learning, and it reports attainment resulting from learning. I hope that this publication will help you cut your way through the thicket of ever-increasing data so that you can find a clear pathway to better ways of using assessment information to make those wise decisions.



Dr Mary Boustead
General Secretary

CONTENTS

1

- 7 INTRODUCTION**
7 Using this publication

2

- 8 WHAT DO WE REALLY MEAN BY ASSESSMENT?**
8 Purposes and kinds of assessment
9 The uses we make of assessment information
9 Accountability and summative assessment
12 Is it true that formative assessment is more valuable than summative assessment?
13 Activities 1, 2, 3 and 4

3

- 16 WHAT DO WE MEAN BY VALIDITY AND RELIABILITY?**
16 Validity
18 Reliability
21 Can we construct an assessment system that is both reliable and valid?
21 Continuous assessment versus testing
22 Using ICT for assessment
23 Activities 5, 6 and 7

4

- 24 WHAT ARE THE BASES OF COMPARISON?**
24 Norm-referencing
25 Are examinations getting easier, or is achievement rising?
25 Criterion-referencing
26 How specific should assessment criteria be?
27 Ipsative-referencing
28 What does all this mean for classroom teachers?
28 Activity 8

5

- 29 HOW CAN WE ANALYSE ASSESSMENT DATA AT CLASSROOM OR SCHOOL LEVEL?**
29 What do you need to know?
29 Ideas for analysis and use of data
30 Progress measures
33 Analysing test scripts
35 Using assessment data to set targets for pupil progress and attainment
37 Pupil involvement in the target-setting and self-assessment process
37 Comparing your pupils' assessment information with pupils in similar schools
37 Activity 9

6

37 HOW DO WE MOVE FROM ANALYSIS TO ACTION?

- 38 Addressing the issues
- 39 Fitting in with other developments
- 39 Planning well
- 39 Monitoring and evaluating
- 39 Worth the time and effort?

7

39 FURTHER INFORMATION

- 40 Useful websites
- 40 Useful books
- 41 Useful videos



ABOUT THE AUTHORS

Sue Swaffield is a lecturer at the University of Cambridge Faculty of Education. She works in the areas of assessment, leadership and school improvement, teaching on Master of Education and other Continuous Professional Development programmes. Before moving into higher education, Sue worked in local authorities as a general adviser with special responsibility for assessment. She has taught in primary and secondary phase schools, in the UK and abroad. Sue was president of the Association of Assessment Inspectors and Advisers (AAIA – now known as the Association for Achievement and Improvement through Assessment) from 1999 – 2001.

Pete Dudley is Project Director for Classroom Learning with the National College for School Leadership Networked Learning Group. Prior to this he was head of School Improvement and Lifelong Learning in the London Borough of Redbridge. He spent many years teaching in primary and secondary schools in East London and abroad before joining Essex LEA as an assessment adviser in 1991. He was a linked school adviser for five years and led the Essex strategies for achievement-raising, teaching, learning and curriculum.



INTRODUCTION

Numbers do provide a useful shorthand way of describing, communicating and measuring what is happening. The challenge is to ensure that educators and the public understand both the possibilities and limitations of such information. Educators and the public need to develop assessment literacy in order to examine student work and performance data of all types and to make critical sense of it. (Earl et al, 2000)¹

There is a growing emphasis on performance as measured by the standards attained by pupils. Teachers are being expected to make more and more use of assessment data. We need to be able to make sense of all this information, in order to improve learning and teaching.

This book aims to assist teachers in developing ‘assessment literacy’ – in other words, an understanding of the issues surrounding assessment and performance data. This understanding relates to assessment in all subject areas, not just in English or literacy. It has been written to help teachers see behind the figures and to know what questions to ask about various forms of assessment data, so that subsequent decisions are well founded.

We should not take assessment data at face value, but nor should we simply dismiss it. The appropriate and thoughtful use of assessment data can have a very positive impact on pupils’ learning and the effectiveness of our teaching. However, assessment data does have limitations, and we need to understand these so that we can make wise decisions, not jumping to inappropriate conclusions, or planning actions on false assumptions.

Having read this book, it is hoped that teachers will be better able to use assessment data to make wise decisions. Specifically, the book aims to assist teachers, parents and school governors to:

- understand assessment processes
- understand the language of assessment
- be able to critique externally generated assessments which are used in schools
- be able to apply their understanding to their internal assessments in order to generate confidence in the results.

Assessment literacy for wise decisions does not attempt to give particular guidance on how to carry out assessment, although the understanding that it seeks to develop should help inform teachers’ assessments. It does not, for example, go into detail about using assessment data for whole school target setting. Rather, it focuses on the teacher in the classroom and on the use of assessment information by subject and/or year group leaders.

Using this publication

The different sections of the book address different aspects of assessment, and so you can use it as a reference guide to clarify particular points. The book can also be read from beginning to end, as the topics are addressed in a logical order.

To help you think about the issues and make wise assessment decisions, activities are featured at the end of each of the main sections. There is also a ‘further information’ section with useful websites and suggestions for further reading on page 39.

WHAT DO WE REALLY MEAN BY ASSESSMENT?

This section:

- the different purposes of assessment
- outlines the differences between formative, summative and evaluative assessment
- discusses the different ways in which assessments are used.

There are few words so likely to provoke heated debate in an educational context as ‘assessment’. In general terms, assessment is what we do when we take stock of how a learner is progressing. How we do this, and why we do it, varies tremendously. In one extreme sense we can say that we make an assessment every time we respond to a pupil’s question. This is because we consider what the question reveals about the pupil’s level of understanding, and then we frame a response to take account of that and to help them make progress.

It can be powerfully argued that as 99 percent of babies accomplish the enormously complex task of learning to talk – purely by making sense of feedback from adults, siblings and contexts – then we should not tolerate pupil underachievement. If a pupil is not making progress, then it is because we are not giving them the right feedback to enable them to learn (or so the argument goes). This is because we may not have made the correct assessments.

In another sense, we can say that assessment is what happens when pupils take formal tests and examinations, or when they are screened for particular learning difficulties.

Purposes and types of assessment

The range of meanings we attach to the word ‘assessment’ is partly why so much controversy surrounds it. Assessment is used to determine how or what we teach a pupil or a group next.

This is *formative* assessment, which happens all the time. We often plan to make some formative assessments in plenary sessions, but we will also make them on an unplanned basis during the course of a lesson. Good teachers are adept at spotting when they have made an off-the-cuff formative assessment and will make a mental note to use the information later. Some teachers devise systems for recording significant formative assessments, but most formative assessments are not recorded. The results – or information – from formative assessments are used to frame feedback to pupils and to plan subsequent lessons.

Assessment is also used to determine how well a pupil or group have grasped something after a period of time – maybe at the end of a teaching unit or course of study, maybe at the end of a key stage or a whole phase of education. This is *summative* assessment. It is often highly formalised – as seen in national curriculum tests and public examinations. It may require teachers to make a judgement about the progress pupils have made, based on their work over time. This happens with formal Teacher Assessments using national curriculum levels. Sometimes summative assessments are made using tools that are more general and not specific to the content of lessons. An example of this would be the way some schools double check against a teacher’s assessment by using commercially available standardised tests at fixed intervals to gauge pupil progress.

Assessment information may be used on its own for a number of critical purposes. These include:

- providing a baseline against which to judge the progress of a pupil or a group
- providing a subsequent measure of the progress of the pupil or group and then making a judgement about how good that progress is.

This can result in an evaluation of what has proved successful (or unsuccessful) with the pupils in terms of teaching and learning. It can also help to identify what should be taught differently next time – or even what needs to be re-taught. This is *evaluative* assessment, because it evaluates the effectiveness of the learning, the teaching and – increasingly – the teacher.

Assessment information may be used evaluatively:

- to judge the effectiveness of a piece of curriculum, a course of study, or particular materials or resources
- to judge the strengths and weaknesses of the teaching, or the performance of a teacher, a department or a school.

Alternatively, assessment information may be used evaluatively to identify where a pupil may be having particular difficulties. In some cases, subsequent assessments may be applied to diagnose a specific learning difficulty or to trigger additional teaching time. These are often referred to as *diagnostic* assessments.

The uses we make of assessment information

Assessment is a ‘hot topic’ and often described as ‘high-stakes’ because it is used as a basis for making decisions which profoundly affect people’s lives. Forty years ago, assessment was widely used to debar pupils at the age of 11 from access to the public examination system, at the age of 16 to deny them the opportunity to study for A-levels, and again at the age of 18 to ensure that only five percent of the

population could study at university. Today, the focus has swung towards identification of those pupils who are likely to fail, and creation of additional support for them in an attempt to counter that early failure.

Examples of this can be seen in the current literacy intervention strategies at Year 1, Year 3, and Year 7. Conversely to ‘catch-up strategies’, there are now moves to ‘fast track’ high-attaining pupils to complete a key stage a year early.

Life chances are greatly affected by assessments made at school. Pupils can be grouped according to assessment information – this can affect their self-esteem as learners, their expectations of success, and also those of their teachers.

Accountability and summative assessment

Summative assessment is increasingly used for accountability purposes. It is used as a measure of the success of a teacher or a school in terms of the progress made by pupils between assessment points. Exam results and tests are used by Ofsted inspectors (Estyn in Wales) to help calculate the ‘value for money’ provided by schools.

As we have seen, assessment can have huge effects on pupils, schools and teachers. We need therefore to understand its potential, and its limitations. We also need to know where and when to use different types of assessment to best effect. Table 1 (overleaf) gives examples of common assessments, and analyses their purpose, possible uses and their limitations. Reference is made throughout the rest of this book to the types of assessment illustrated in this table.

TABLE 1

Common types of assessment in use in schools

Activity	Timescale	Assessment purpose	What could these assessment outcomes be used for?	What could these assessment outcomes not be used for?
1 A pupil's question/comment reveals understanding which is either greater or less than the teacher had predicted. The teacher has to attune her/his response carefully and make a mental note of the exchange.	Ongoing	Formative	Planning immediate subsequent teaching, or redirecting learning.	Making major decisions about longer-term courses of study.
2 A teacher marks a pupil's work or judges her/his oral contribution to a plenary session. S/he feeds back on where the educational objective has been met, where there are gaps, and what the pupil can work on next in order to improve.	Daily/weekly	Formative	Planning future learning – especially by the pupil. Checking that planned learning has broadly been achieved, and recording major issues for future teaching.	Ranking or finely grading the work – because the judgements are about next steps.
3 A teacher plans a focused 'assessment lesson' connected with a series of lessons, where pupils apply their recent learning to a new context (e.g. after half a term on separation methods in science, they are asked to solve a murder mystery using forensic techniques which demand appropriate use of different separation methods).	After a series of taught sessions and around three-quarters of the way through a unit	Formative Summative	Feeding back to pupils on what progress they have made and what to target next. Planning subsequent learning strategies. Checking that pupils have grasped the learning and can apply it. Planning subsequent teaching as a result. Judging the progress they have made. Judging the national curriculum level that pupils are operating at.	Making major decisions about longer-term courses of study.

Activity	Timescale	Assessment purpose	What could these assessment outcomes be used for?	What could these assessment outcomes not be used for?
<p>4 A teacher plans a session where pupils respond to questions (orally or in writing) on a recent topic or module of work.</p> <p>A teacher marks all the pupils' responses, and finds that many have misunderstood a key concept which will need to be re-taught.</p>	After a series of lessons, at the end of a unit	<p>Summative</p> <p>Evaluative</p>	<p>Checking levels of understanding. Judging the national curriculum level the pupils are operating at.</p> <p>Deciding whether pupils are on course to succeed in the longer term. Broadly judging how successfully the unit was structured or taught.</p>	Planning subsequent learning strategies in the topic to rectify misunderstandings (because it is too late, the unit is finished. The teacher may adjust a future unit, or repeat its content another time).
<p>5 Pupils sit a controlled multiple-choice reading test.</p>	Six-monthly	<p>Summative</p> <p>Diagnostic</p>	<p>Checking to see if the score has increased since the test was last taken by the same amount as most pupils of the same age.</p> <p>Where a score is worryingly out of line, applying a further test or assessment to diagnose a possible problem.</p>	Planning subsequent learning strategies or evaluating how successfully the unit was taught or structured (because in most cases, the information generated is not sufficiently diagnostic).
<p>6 Pupils sit a national curriculum mathematics test which samples pupils' abilities to use taught skills.</p>	Annually or bi-annually	<p>Summative</p> <p>Predictive</p>	<p>Giving pupils a national curriculum test-level and age-related score.</p> <p>Broadly categorising pupils for the next school/key stage, in order to inform planning or grouping.</p> <p>Setting targets for the group to achieve by the end of the following year or key stage.</p>	Planning subsequent learning strategies, unless a detailed test item analysis is undertaken (see page 33: <i>Analysing test scripts</i>).

Is it true that formative assessment is more valuable than summative assessment?

Table 1 (page 10) shows that it is possible for formative assessments, linked to high quality feedback, to be used to help pupils plan strategies for subsequent learning. This has led to the expression 'Assessment for Learning' (AfL). This concept is defined by the Assessment Reform Group as:

'the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there'.

(Assessment Reform Group, 2002)²

Pupils who are used to using AfL become better at learning independently and can make increased progress. AfL also shows that formative assessments and summative assessments which are related to the precise skills or content of the teaching (like the third example in table 1 on page 0), can be used to give feedback to help pupils target future learning. In addition, these types of assessment help teachers to judge both the effectiveness of their teaching and the progress pupils have made. It also helps them to plan the next phase of teaching. This can therefore be viewed as assessment for teaching. Pupils can agree their own next-step learning targets on the basis of the outcomes of the assessments.

It is also clear that assessments which are made by methods that are divorced from (or at a time that is divorced from) the direct teaching and learning that has been taking place in the classroom (such as de-contextualised commercial tests) have much more limited value in their subsequent use for teaching and learning.

In answer to the question 'which is more valuable?' neither approach is necessarily more valuable than the other. But where formative assessment is made in a context where the teacher knows the level of demand an activity is making on the pupils (for instance, the national curriculum level at which the science separation techniques have been pitched), then formative assessment activities can also be used to make summative assessments as well.

The teacher in the third example in table 1 (page 10) is able to do this and, if necessary, make vital adjustments to teaching for the remainder of the unit. Pupils can also be shown where they need to improve. The teacher in the fourth example cannot do this. If the unit is at the end of a year or at the end of a key stage, then the opportunity to revisit or re-teach a unit may have been lost. It is seldom possible to do the reverse and make formative judgements from summative assessment data.

What must be remembered is that pupils should be the ultimate beneficiaries of assessment. Where pupils are clear about what counts as progress and success, how assessments are made, and how they can use assessment feedback, then they can become more effective learners.

'As teachers bring students into the assessment equation, thus demystifying the meaning of success in the classroom, they acknowledge that students use assessment results to make the decisions that ultimately will determine if school does or does not work for them. Our collective classroom assessment challenge is to be sure students have the information they need, in a form they understand, and in time to use it effectively.'

(Stiggins, 1994)³

ACTIVITY 1

Use the following questionnaire with colleagues to help establish a shared view of what should be and what is happening in your school/department in terms of assessment.

How important/relevant is the statement for your establishment?					Fundamental Statement	How does your current practice match the statement?				
4	3	2	1			4	3	2	1	
4	3	2	1			4	3	2	1	
1	2	3	4			1	2	3	4	
					Assessment offers all pupils an opportunity to show what they know, understand and can do					
					Assessment practice helps pupils to understand what they can do and where they need to develop further					
					The key learning outcomes of each subject or learning experience (early years) have been identified so that assessments made against these can be used to help develop children's learning					
					Assessments are not restricted to national curriculum subjects					
					Sharing of learning intentions is routine practice, which enables the pupils to understand their role in the lessons					
					Assessment practice in the school enhances the learning process					
					Assessments made by the teachers inform daily and weekly planning and allow learning to be matched to the needs of the pupils					
					Assessment of pupils' learning is reported to parents in a way which identifies achievements and what the child needs to do to improve					
					Pupils are involved in assessing their own work and that of their peers					
					Pupils and teachers work together identifying targets for learning and ways of achieving these					
					Core assessment data on each child is updated each year and passed to the receiving teacher or school to aid future planning					

ACTIVITY 2

List all the assessments that are undertaken in your school/department. Categorise each as formative, summative or evaluative, and note what use you make of the outcomes. (Use Table 1 on page 10 as a prompt, if this helps.)

ACTIVITY 3

What kinds of assessments do you use in your school and how useful are they?

With a colleague complete the following assessment analysis and reflect on any issues the activity raises for assessment policy in the school/department.

Type of assessment/test used	Yes/No	How often?	Who sets and marks the tests?	How is the information used with the pupils?	To whom are the results communicated?	How is the information used to: inform planning, evaluate teaching or other uses?	What issues does this raise for the school?
Teacher assessment							
Annotated plans							
End of unit assessments/tests							
Regular reading tests							
Regular standardised maths tests							
CAT tests							
Other standardised tests (list)							
Pupil self assessment/target setting							
Objectives completed tick sheets							
End of year tests (e.g. QCA Online systems (e.g. Goal)							
Others							

Actions points emerging from this analysis

ACTIVITY 4

How do your views of what should be happening (activity 1) compare with what is *actually* happening (activity 2)? Make a list of the implications of your findings, and plan any necessary actions.

NOTES

WHAT DO WE MEAN BY VALIDITY AND RELIABILITY?

This section:

- outlines the concepts of validity and reliability
- considers how validity and reliability are related
- considers the use of ICT for assessment in relation to validity and reliability.

Whatever the particular purpose of any assessment, there are likely to be consequences for the pupils who were assessed, and for those who use the information. We therefore need to be confident that the outcomes of assessment are dependable. The two terms which are used in relation to the confidence we can have in any assessment are *validity* and *reliability*.

Valid and *reliable* are words which are used in everyday language. Since they have specific meanings for assessment, it is important to understand their more technical usage.

Validity

In assessment terms, validity is taken to be the extent to which any assessment succeeds in measuring that which it originally set out to measure. Although this may sound straightforward and obvious, there are many instances where assessments are not valid. For example, a written test in science may be attempting to assess pupils' understanding of scientific concepts, but it may actually be more of an assessment of their reading and writing skills. For pupils with difficulties in reading and writing, this form of assessment may invalidate any judgements inferred about their learning of the science topic.

Many of the special arrangements that can be put into place for the end-of-key-stage tests are designed to overcome a particular invalidity caused by a pupil's particular circumstances. Pupils who have difficulty

writing, for example, would be unable to demonstrate their mathematical understanding in a conventional pencil and paper test. However, the use of an amanuensis (someone who writes down the pupil's answers) enables the test to assess what it is supposed to assess – the pupil's aptitude in mathematics, rather than their aptitude in writing.

Beyond the broad definition of assessing what it sets out to assess, validity is a complex issue. A number of different types of validity have been identified; we will be concentrating on four. These are *construct validity*, *content validity*, *predictive validity* and *consequential validity*.

Construct validity refers to the extent to which any test assesses the particular 'construct', or underlying skills or attributes, which it is supposed to assess. For example, when a test produces a 'reading' result for a pupil, what exactly is that test giving us information about? Is it: the pupil's ability to read aloud accurately; the pupil's ability to respond in writing to written questions about something that has been read; or is it to do with the pupil's response to the text?

At key stage 1, pupils who are awarded a national curriculum level 1 in reading as a result of their performance on the reading task will have been given the opportunity to 'demonstrate their ability to read aloud from a text, show what they have understood, and give a personal response' (QCA, 2000)⁵. Pupils who are awarded a national curriculum level 3 in

reading as a result of their performance on the level 3 reading test will have been assessed on their ability to read independently and respond in writing to comprehension questions on both a story and information text. Pupils awarded a level 2 in reading will have been assessed through both the reading task and reading test, and so will have demonstrated attainment across a broad range of reading skills.

Content validity is related to construct validity and is essentially about coverage of a scheme of work. For example, does the end-of-key-stage science test assess a representative sample of the knowledge, skills and concepts covered by the relevant scheme of work? Any test which concentrated on the 'Life process and living things' aspects, or which included topics which are not covered until the next key stage, would not have content validity because it is not representative of the curriculum that is being assessed.

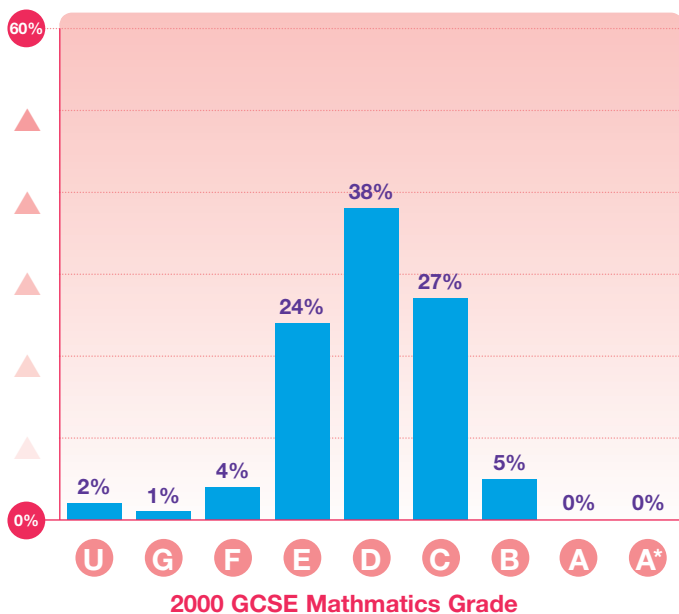
Predictive validity refers to the extent to which the results of one assessment accurately predict performance in a future assessment. For example, are a pupil's A-level results a good predictor of how well s/he will perform in higher education? Are Year 7 CAT scores a good predictor of performance at GCSE? This aspect of validity has become of great interest to the Department for Education and Skills (DfES) as they seek to make links between performance at the end of one key stage and the next.

Great care must be taken when considering the application of the predictive validity of any test, however. It may be that the majority of pupils who attain level 2 at key stage 1 go on to attain level 4 at key stage 2, but this is the likelihood in relation to a whole group, not a certainty for a particular pupil. For any individual pupil, a level 2 at key stage 1 may be followed by any one of a range of levels at key stage 2. There are great dangers in suggesting to any pupil that their future performance is *determined* by their past performance. Devices such as the progress charts within the DfES' Pupil Achievement Tracker interactive software package which in 2003 replaced the former paper based 'Autumn Package' (in England) can be used to demonstrate to individual pupils the range of outcomes that previous pupils have attained (and so open up the possibilities of attainment to them) rather than suggest a pre-determined outcome. For example, as table 2 shows, the most likely outcome for pupils with a key stage 3 average points score of 31 to 33 in 1988 was a grade D at GCSE mathematics in 2000. However, over half of the pupils were awarded either a grade E or a grade C, and five percent (or one in 20), of pupils with the same average points score actually gained a grade B⁶.



TABLE 2

Key stage 3 average points score ≤ 33



DfES, Autumn Package 2000 GCSE/GNVQ, 2000.

It should also be remembered that a correlation between two factors does not necessarily mean that there is causation.

Consequential validity is an idea that has been developed by Messick (1989)⁷ to pull together a number of other ideas about validity. It focuses on the consequences of test results – in other words, the educational and social implications of the way assessments are interpreted and used. This means that even a well-constructed test is not valid if the results are used inappropriately – which moves the idea of validity on from something which is the concern of the test writers to something which is the responsibility of everyone who interprets and uses assessment results.

Reliability

If validity is about the extent to which a test measures what it is designed to measure, then *reliability* is the extent to which you can use the test on different occasions and with different pupils, and still be sure that the same things are being measured in the same way and to the same extent.

A reliable assessment will usually provide results which can be used either:

- to make comparisons with the results of other pupils who have been assessed in the same way, or
- the same pupil's results from an earlier occasion upon which they were assessed. The earlier assessment may have predicted what their result might be now.

Assessments that are designed to provide reliability more than validity are usually termed 'tests'. Test developers sometimes provide an indication of how much a user can rely on the result through such features as 'confidence bands' (see page 20).

Table 3 opposite contrasts one reading assessment which is especially designed to produce highly valid results with another, which is focused more on producing *reliable* results.

TABLE 3

Reliable and valid reading assessments strengths and weaknesses

	Assessing reading – validity model	Assessing reading – reliability model
	<p>The teacher meets with the pupil each month to discuss her/his progress in reading since the last meeting. The pupil's reading log is referred to and the teacher makes notes about the level, range and volume of the pupil's reading, as well as about aspects of the pupil's responses to the reading. The teacher listens to the pupil read and discusses some parts of the text with him or her to ascertain the depth and level of their understanding and use of skills such as inference, use of knowledge of other texts in similar genre to make predictions, understanding of characterisation, and plot devices. The teacher and pupil also discuss other works by the same author. The discussion then goes on to cover parental/peer support for reading, and a list of possible titles for the pupil to tackle over the coming month with some specific skills-related targets. Finally, both teacher and pupil discuss national curriculum level-related criteria, and agree where the pupil has made progress and where the focus for the next stage of the pupil's reading development will be. The teacher adds these notes to the reading assessment profile, and makes a judgement about engagement and progress.</p>	<p>The teacher gives out a set of commercially-available reading tests. The pupils sit the unseen tests in silence for a set period of time. The tests demand that pupils read a passage of text and then respond to items requiring them to have understood aspects of it. Items specifically relate to: the lexis used; how meaning was conveyed through the syntax and grammatical devices in the text; what deductions could and could not be made from parts of the text; what could and could not be inferred about character and motive in parts of the text; and how the extract resembled a particular genre.</p> <p>The items nearly always require a one-word answer or offer a multiple choice of four possible answers, one of which is correct.</p> <p>The tests can be marked by anyone applying the same rules. They produce a raw score result for each pupil, which can then be compared with the average scores of other pupils whose ages fall within the same calendar month as the pupil. A comparable 'age-related standardised score' can then be generated.</p>
Outcomes	Both teacher and pupil have a shared, in-depth understanding of the progress which is being made by the individual pupil across a wide range of reading skills, including understanding, motivation and behaviour. Next steps to ensure good progress are also established.	The teacher has a raw score result: a number which can be compared to the number generated the last time the pupil took the test to find out whether the pupil's standardised score has gone up or down. This comparison could be used to establish whether or not the pupil is making average progress, and whether their progress rate has changed.
Strengths	The results of this assessment offer lots of information about the pupil and her/his reading, as well as the next steps they should take. The assessment gives a strong sense of the general rate of progress that the pupil is making, as well as indicating formatively what the next steps should be. The pupil is closely involved in the assessment, and benefits from it as a learning experience.	The results of this assessment are designed for comparison, and they provide good comparisons with data from large numbers (1,000 pupils in each standardisation sample, for instance). If the text in the test is in any way related to what has been the focus of learning, then the test can give good information about the success of teaching techniques. This may inform a teacher how to teach that item another time. Results can also be used predictively.
Weaknesses	<p>Because the information is largely in words rather than numbers, it is hard to use the results to make comparisons. It is hard to process the data statistically – in order to present it in other formats or to compare it with other assessment data, for example.</p> <p>The test requires a good deal of time and considerable teacher skill. The teacher must have good subject knowledge and know the national curriculum levels well. The teacher must also act in a way that is consistent with other teachers and assessors. It would be very difficult for a teacher who did not know the pupil, or who had not assessed the pupil before, to produce accurately referenced assessment information or to make as informed a judgement as is possible about her/his progress.</p>	All too often, the material in the tests does not relate directly to what has been taught. Consequently, the outcomes detailed above will not occur. These assessments are easily rendered invalid: for example, if a pupil has done the test before and remembers the answers. If the text is demotivating in content to any specific pupil, this may affect her/his performance, and if a pupil is worried by the conditions of the test and finds it difficult to judge precisely what the items are requiring, then the tests can become a test of a pupil's ability to do such tests as much as a test of their understanding and response to the text in the test. Multiple-choice items can confuse a reader who understood the text perfectly but whose judgement was affected by the similarity of the choices offered – not by the text.

It can be seen then that there is usually a trade-off between validity and reliability. There are lots of examples of test 'howlers': these are often the result of a pupil failing to grasp the intention of a question. One example is a mathematics test item for 14-year-olds that provided a picture of a ladybird with equal numbers of spots on each wing. Under this picture was a series of questions about even numbers and matching numbers. In answer to the question 'Could a ladybird have 21 spots?', one pupil – looking for a more theo-philosophical basis to the question than was required – answered 'Yes, because if God wanted the ladybird to have 21 spots, then God could give the ladybird 21 spots'.

Clearly the attempt to make the question more valid by basing it on an illustration (the ladybird) misfired! But when you think about it, the question 'Could a ladybird have 21 spots?' is hardly an everyday question.

It is important to remember that it is the test itself which is 'reliable', not the assessment of the individual pupil. If information from a test is to be used to make decisions wisely, then the reliability of that test must be taken into account. Many teachers are not aware of the lack of precision in a seductively 'precise-looking' standardised score of, say, 110. It is worth remembering that a score of 110 on a standardised test means only that there was a 90% probability that the pupil would score between 102 and 118 (110 being used because it is the midpoint). Not really very precise at all!

The reliability of a test may be given through confidence bands or a reliability coefficient. Always check the reliability details supplied with any standardised test being used. Confidence bands provide a visual representation, while a reliability coefficient provides a numerical representation. An example of how confidence bands work is given in the box below.

CONFIDENCE BANDS

- As the standardised scores (from the level threshold) tables are derived only from one short test, some margin of error is inevitable. To indicate how wide this margin of error is likely to be, a '90 percent confidence band' is calculated. This means that you can have 90 percent certainty that the child's true score lies within this confidence band.
- For the reading test, the 90 percent confidence band is plus or minus 8. So, for example, if a child has a standardised score of 110, you can be 90 percent certain that the true score is between 102 and 118.
- For the spelling test, the 90 percent confidence band is plus or minus 8. So, for example, if a child has a standardised score of 110, you can be 90 percent certain that the true score is between 102 and 118.
- For mathematics test, A and test B, the 90 percent confidence band is plus or minus 5. So, for example, if a child has a standardised score of 110, you can be 90 percent certain that the true score is between 105 and 115.
- For the mental arithmetic test, the 90 percent confidence band is plus or minus 9. So, for example, if a child has a standardised score of 110, you can be 90 percent certain that the true score is between 110 and 119.

Qualifications and Curriculum Authority 2000⁸

Reliability coefficients are expressed as decimals (e.g., $r = 0.9$), with the reliability increasing as the figure becomes closer to one. A reliability coefficient of 0.85 to 0.9 is commonly regarded as being high enough to be acceptable, but even then ‘...the errors in pupils’ scores that are implied may mean that a significant proportion are given the wrong grade’ (Black 1998)⁹. Anything with a reliability coefficient of less than 0.85 should be treated with a great degree of caution.

Can we construct an assessment system which is both reliable and valid?

National curriculum tests and public examinations have struggled with ensuring both reliability and validity for years. Coursework and Teacher Assessment typify the attempts to ensure greater validity. Externally marked tests and mark schemes represent the focus on reliability.

Continuous assessment versus testing

In the 1970s, continuous assessment began to become more commonplace. In the past it had been most frequently used in vocational training – in apprenticeships, for instance. Proponents of continuous assessment argued that it presented a more balanced picture of a learner’s level of performance than did a one-off written examination. Continuous assessment, it was argued, also allowed better, more valid assessment of skills and application of knowledge, whereas formal examinations frequently tested recall and writing skills, and were also open to all the arguments about validity that have been considered earlier in this section.

In 1972, the school-leaving age was raised to 16 years, meaning that almost all pupils stayed at school to take some formal qualification at that age. The Certificate of Secondary Education (CSE), which had been devised by secondary modern school teachers in order to acknowledge the achievements of pupils who would otherwise have left school with no formal recognition of their achievement, was being taken by an increased number of pupils. Many CSE courses were partly assessed by coursework that was completed during the fourth and fifth years (now Years 10 and 11). The need to submit coursework of a high quality in order to achieve the certificate also acted as an intrinsic motivator for pupils. A grade one CSE became accepted as the equivalent in status to a GCSE O-level pass (what would now be a GCSE grade C).

During the late 70s and the 80s, unemployment was rising and one of the main extrinsic motivators for pupils to work at school – that they would stand a better chance of getting a job when they left at 16 – was fast disappearing. Intrinsic motivators became increasingly necessary. Schools introduced Records of Achievement to celebrate success. Both GCEs, CSEs and later (from 1986) GCSEs contained increased levels of continuously assessed work up to 100 percent. Some CSEs, known as Mode 3s, were set and assessed at school level, with both the coursework and the marking being validated by the exam board.

The debates around the merits of continuous assessment – as opposed to examinations and testing – continue today. Many argue that the rise of female achievement, particularly in some subjects, reflects the fact that continuously assessed courses favour female work styles (statistically, girls do better in coursework and boys do better in examinations) far more than it reflects any successes of educators in overcoming sexism in the curriculum. In the early 90s, the Government ruled that coursework opened the door to unfairness, with some pupils benefiting from more support than others. The percentage of coursework in any GCSE syllabus was then severely restricted.

Teachers of some subjects – notably English – have tended to favour continuous assessment, while others continue largely to prefer tests. Many have held what could be seen as self-contradictory positions, arguing on the one hand that continuous assessment in the form of national curriculum teacher assessment is unhelpful, while simultaneously arguing passionately for increases in permitted GCSE coursework.

Research has shown that where teacher assessment is well moderated and teacher subject knowledge is sufficiently developed to ensure high quality assessment, then such teacher assessments can be more reliable than externally marked national curriculum tests (William, 2000)¹⁰. This is because it can balance out ‘off-days’, and represent more readily what the pupil knows, understands and can do. To ensure that teacher assessment is reliable, as well as valid, it is essential that teachers’ work is moderated, both with other teachers in the school and also from beyond school, using both local and national exemplars of attainment in pupils’ work. Taking part in moderation activities is also good for teachers’ professional development, and has many other benefits such as teachers becoming clearer about assessment criteria and how to interpret them, and teachers’ confidence in information transferred between schools being improved.

A significant question is raised by the use of ‘cut-off points’ in public tests and examinations. In some GCSE mark schemes, only 15 marks separate an A and a D grade, and yet 15 is the standard margin of error. There is little evidence to date of litigation around the reliability or validity of public assessment systems in the UK, except in the realm of special educational needs. However, with the life chances of pupils and even the remuneration of teachers so tied to pupil performance in such tests, this is sure to change. The return of GCSE and A/S and A2 examination scripts to schools may hasten an increase in disputes over marking.

Using ICT for assessment

Information and communications technology (ICT) is helping to bring validity and reliability together at a classroom level through well-designed tests which are marked electronically. Rapid comparisons of results from large databases are also made possible by this approach. For instance, software being developed at the Centre for Research on Evaluation, Standards and Student Testing at the University of California in Los Angeles (UCLA), requires pupils to solve problems during and at the end of units of work. The UCLA computer-based system has records of hundreds of learners (similar prior attainment, age, preferred learning style, same ethnicity, same gender, etc) and so each pupil’s results can be compared with those of his or her peers, as well as with a standard. The software also provides information about the pathways that the pupils took to reach their solutions, and so it provides information that can be used both formatively and summatively.

Another technique using ICT that can be very powerful, both formatively and summatively, is *concept mapping*. This is used both at the beginning and the end of a unit of work. Many teachers use concept mapping to elicit from pupils their mental conceptual map of the relationships between different elements of, say, scientific understanding. These maps can help teachers to plan teaching to match the areas where the maps show weak or incorrect understanding or conceptualisation. The post-teaching map can reveal the degree of progress that the pupil has made.

ICT-based testing systems such as 'GOAL' are also attracting a lot of interest. This is understandable, as they are heavily promoted and claim to save that valuable commodity – teachers' time. ICT enables the results of pupils taking Qualification and Curriculum Authority (QCA)-style tests to be compared rapidly with those of similar learners on the scheme's database. The similarity of the GOAL tests to the QCA tests enables better predictions and preparation for end-of-key-stage national curriculum tests.

There are several dangers associated with ICT use in assessment, however. Firstly, the very nature of the ICT-based tests means that responses are almost always either multiple choice or a one-word answer. This restricts the range of learning that can be assessed, and leads to a narrow view of what constitutes valued learning. Secondly, an over-reliance on ICT-based tests can reduce a teacher's role in assessing the knowledge, skills and understanding which constitute learning at a particular level. In the long term, this leads to a de-skilling of teachers. Thirdly, ICT based tests can result in the feeling that *'It's alright, we've got assessment sorted'* which ignores the most important aspects of a formative assessment which supports learning.

Of course ICT, used wisely, can yield useful (if limited) information, and free the teacher to make best use of the information gained and to concentrate upon assessment for learning in the classroom. In summary, ICT-based approaches to assessment are generally high on reliability, in that the scoring of answers is reliable, but low on validity because they are only able to test a narrow field of learning.

As with any system which defines its measures and then sets them as goals, we are in danger of 'measure fixation' (Kellner, 1997)¹¹. Measure fixation can develop where the measure and goal become so inextricably linked that the measure becomes an end in itself – the validity (or reliability) of the tests becomes of secondary importance. This can potentially lead to an entire system seriously losing the plot! Kellner draws a parallel with the Thatcher government's fixation on money as both the measure and outcome of economic policy, and the long-term effects on the economy of not looking more regularly at a wider range of indicators of economic performance.

Although it might seem that the 'best' assessment will have both high validity and high reliability, we have seen that, due to the tension between the two, there is likely to be some trade-off between validity and reliability. In using assessment wisely, we would do well to follow Wynne Harlen's advice that an assessment should possess both high validity and optimum reliability (Harlen, 1994)¹².

ACTIVITY 5

Look at the tests that are used in your school/department. What information is given about their reliability? What does this say about how you should use the outcomes of these tests?

ACTIVITY 6

Consider the various assessments used in your school/department. Do they give you information about the things you intended them to, and that you value?

ACTIVITY 7

How can you increase the validity of, and achieve optimum reliability in, assessment in your school/department?

4

WHAT ARE THE BASES OF COMPARISON?

This section:

- explains norm, criterion and ipsative-referencing
- considers the implications of each of these approaches.

Whenever we make an assessment, we are making judgements about performance *in relation to something*. The different ‘yardsticks’ used are essentially other people (norm-referencing), a specified standard (criterion-referencing), and the individual’s previous performance (ipsative-referencing). In this section, each of these different approaches are considered separately, although distinctions between them are not as completely clear-cut as may first appear.

Norm-referencing

Norm-referencing is the approach taken when pupils are compared with other pupils, or with ‘the average pupil’. This was the traditional form of assessment experienced by many of today’s parents, when pupils were judged to be ‘fourth in the class’ or whatever. Norm-referencing responds to requests from parents to know how their child is performing in relation to the rest of their classmates.

Norm-referencing within the classroom tells us whether a pupil can or cannot perform better than other pupils, but it does not tell us anything about what that individual can or cannot actually do. The information is only comparative, and so could be useful if competition with others is the purpose, such as for selection by ranking. Norm-referenced assessment says nothing about what the pupil specifically knows, understands and is able to do. Similarly it does not tell us about any particular difficulties, or what the appropriate next steps in learning are for the individual pupil.

The notion of an ‘average’ pupil, or a ‘norm’ of performance, leads to a classifying of individual pupils as ‘above’ or ‘below’ the norm and of them as having ‘passed’ or ‘failed’. Norm-referencing and its associated ideas of ‘above and below average’ are closely linked to concepts of fixed intelligence, and of high and low ability. As we come to understand more about the nature of learning, the functioning of the brain, and the importance of expectations and self-belief, it is clear that a norm-referenced approach to assessment does little if anything to assist learning, even if it may be used for labelling in different ways.

When tests are norm-referenced, it means that the results are expressed in relation to the results of a whole group. In order to establish the information against which an individual’s results are compared, a sample is used. This sample has to be representative of the whole population, and, generally speaking, the larger the group, the more meaningful the result. The results from the sample are scaled to give a normal bell-shaped curve, which is used as the basis of comparison for any individual result, and enables statements such as *‘the pupil scored within the top 15 percent’* to be made.

Standardised tests are examples of norm-referenced tests. When the tests are being developed, they are taken by a large representative sample of children of particular ages. The results of each child can then be compared with the ‘national average’ for that age. Reading ages are obtained from norm-referenced tests that are standardised in a particular way: in relation to age in years and months. When using standardised tests, be aware of how long ago the norming was

done, and the size and representativeness of the sample. If the tests have not been re-standardised for some time, then the 'average' with which individuals are compared to is no longer representative of performance in schools.

Are examinations getting easier, or is achievement rising?

Norm-referenced tests, with their in-built comparison, can lead to such apparent absurdities as 'the majority of pupils are above average'. Explanations for this may lie in the original sample (was it large enough, and really representative?), and from changes over time (when was the test last standardised?). Teachers can become familiar with the form of a test, and so be able to prepare their pupils better. Pupils overall may indeed get better – raising standards is the main thrust of current Government policy and school practice.

Despite having grade-related criteria, A-levels (A/S and A2) and GCSEs are basically norm-referenced, in that the percentage of pupils achieving any particular grade is not allowed to vary very much from one year to another without the scrutiny of senior examination board officials. This can lead to confusions over what changing results from year to year mean in relation to raising standards. How should a greater proportion of grade 'A's be interpreted? This question seems to tax the media every summer, and all too often this leads to a belittling of the hard-earned achievements of pupils and teachers.

Actual improved performance by pupils would be reflected in a greater proportion of grade 'A's, provided that there were no statistical adjustments (for example, to produce comparable proportions of pupils attaining each grade).

Accusations that 'exams are getting easier' have led to studies being commissioned to examine the situation. In 1996 the School Curriculum and Assessment Authority (SCAA) published a report on '*Standards in Public Examinations 1975-1995*'¹³ that looked at English, mathematics and chemistry at the ages of 16+ and 18+. The conclusions were that the nature of the demand of the examinations had changed over time, but that, on balance, the level of demand on pupils was broadly similar.

The norm-referencing underpinning the allocation of grades in public examinations also makes any comparisons between different subjects problematic. This is particularly the case when different pupils study different subjects. A grade 'A' in any subject indicates that the pupil is in the top 'x' percentage of students who took that subject; other subjects may have been taken by completely different groups of pupils.

It is much easier to see clearly whether standards are rising or not if a different form of referencing is used – that of criterion-referencing.

Criterion-referencing

Criterion-referenced assessment allows judgements to be made about a pupil's attainment against pre-specified criteria, irrespective of the performance of other pupils. Obviously, when setting that pre-specified criteria, some consideration of what is appropriate, or can be reasonably expected, has to take place. Notions of norm-referencing therefore underlie much criterion-referencing. For example, the 'expected' national curriculum level of attainment for an 11-year-old is level 4.

If they have been well designed, criterion-referenced assessments show what each individual knows, understands and can do. By implication, the next steps for the pupil are those criteria not yet grasped, or in the case of the national curriculum, criteria at the next level.

Just as with the norm-referenced public examinations at the ages of 16+ and 18+, there have been concerns expressed about the comparability of the criteria-referenced national curriculum tests over time. The comparability of key stage 2 reading tests was challenged by research undertaken by Mary Hilton (Hendry, 2001)¹⁴. In 1999 the Secretary of State instructed an independent panel chaired by Jim Rose ‘to consider the Qualification and Curriculum Authority’s (QCA’s) arrangements for setting and maintaining the standard of key stage 2 tests for English and mathematics with a view to reinforcing the credibility and integrity of the tests for this and future years’. Their conclusion was that ‘the public, schools and parents can be confident that the marks required to achieve level 4 on the 1999 English and mathematics tests reflect the nationally-expected standards for most 11-years-olds studying the national curriculum, and are comparable with earlier years’ (Rose, 1999)¹⁵. Ultimately, as with any testing, it is impossible to have absolute comparability year on year.

How specific should assessment criteria be?

One of the major tensions with criterion-referenced assessment is just how specific to be. High specificity and a great deal of clarity are very helpful for designing appropriate learning, teaching and assessment activities, but there is a danger that teaching can become so tightly focused that any unanticipated opportunities for learning are not grasped. The criterion-referenced route can also lead to a tendency to overcomplicate and to break down the focus of assessment into very small, separate

parts. This is often referred to as ‘atomistic’ assessment, and was the case with the original statements of attainment when the national curriculum was first introduced.

The move to level descriptions changed Teacher Assessment from being an assessment of pupils’ attainment against separate criteria, resulting in a great many ‘can do’ statements which were aggregated using a formula to produce a single level, to a ‘best fit’ judgement against a number of criteria which are drawn together into a prose statement. Aggregation, by whatever method, pulls together judgements about performance on a number of criteria into a single manageable level grade, but in so doing loses all the richness of information which went into making that assessment. A national curriculum level gives only a very broad knowledge of a pupil’s performance, and does not provide specific information about any strengths and weaknesses. The same Teacher Assessment level could be correctly given for a pupil who demonstrates all aspects of a particular level description, for another pupil who is working at the level above in some aspects but at the level below in others, and for a third pupil who also demonstrates a range of performance but in different aspects of the level description. Similarly, in tests and examinations where marks for different elements are totalled, the same final score could represent completely different profiles of performance, since weaknesses in some areas will be balanced out by strengths in others.

As teachers’ subject knowledge of what distinguishes each level and their skill at assessing pupils’ performance have grown over time, so has the ability to split the national curriculum levels into three (c, b, and a). The reasoning may go something like ‘Taken overall, level 5 is best fit level, but this pupil does show elements of attainment at level 6, so for internal monitoring my assessment is level 5a’. However, since the levels within a grade are related to the

closeness of the best fit, a grade within a level provides a guide to a pupil's attainment, but no specific information about what that pupil can or cannot do. In contrast, the 'P-scales' (DfES 2001)¹⁶ are criterion-referenced. P-scales (which are intended for use with pupils with special educational needs and allow small steps of progress to be recognised) subdivide levels 1 and 2 in English and mathematics, and provide descriptions of eight stages leading to level 1.

Summary statistics are often used to compare individual pupils or schools. In doing so, it is important to remember that any single score or level could have been arrived at from a wide variety of individual judgements, and so a level or grade gives no specific information about a pupil's performance. Much more information is needed if teachers in the next year group or school are to build upon pupils' prior attainment.

Another issue related to single levels for subjects (and this is often a cause for dispute between teachers) is the comparability of levels between subjects, and even more contentiously, the comparability of levels in the same subject between different key stages. These issues are considered in Dylan William's '*Level best? Levels of attainment in national curriculum assessment*', a publication produced by the Association of Teachers and Lecturers in 2001¹⁷. Research has shown that where pupil ages and identities are hidden and teachers focus on processes and skills alone (*writing, application of mathematics and scientific enquiry*), there is considerable agreement about performance by teachers from both key stage 2 and key stage 3 (Essex County Council, 1999)¹⁸.

One of the main features of criteria-referenced assessments is that they can be designed so that the majority of pupils are able to answer most of the questions. Norm-referenced assessment aims to spread the pupils out in terms of their performance, and so some items which only a few pupils are expected to tackle are included.

In the UK, pupils generally are taught in groups of similar age, and move with their cohort, whatever their attainment. In some countries – France, Germany and Hungary for example – pupils can only progress onto the next year in school if they are successful in criterion-referenced tests. Pupils who do not 'make the grade' must retake the year, typically being taught alongside younger pupils, with all the attendant social and self-esteem implications. There are also similar concerns to be taken into account if the advancement of high-attaining pupils into older age groups is being considered.

Ipsative-referencing

Although the word may not be familiar, teachers use ipsative-referenced assessment every day. In Latin, 'ipse' means 'self', and ipsative-referenced assessment is when a pupil's performance is measured against that same pupil's previous performance. This is what lies behind a teacher saying to one pupil 'I'm really pleased with this work' but if presented with exactly the same from another pupil would potentially respond with chides and directions to improve.

Ipsative-referenced assessment can only work if the teacher has detailed knowledge of the pupil's previous performance, but many people feel it to be the most authentic form of assessment. Ipsative-referenced assessment is applicable to and valuable for pupils at either end of the attainment range, as well as all the pupils who are in between it. After all, it is by improving on our previous best, whatever that

might be, that each of us learn and progress. It is ipsative-assessment that allows challenging and appropriate targets to be set, 'within a pupil's extended grasp'. It is often the ipsative-referenced aspects of reports and consultations that parents value most.

What does all this mean for classroom teachers?

Criterion-referenced and ipsative-referenced assessments are the assessments that support learning, and should therefore be staple fare for teachers. There can be little justification for norm-referenced assessment unless direct comparisons with others are really necessary – this is rarely the case. We need to be aware of the various forms of referencing which lie behind different assessment strategies and tools, and not to read more into any reported results than they actually provide.

ACTIVITY 8

Think of a pupil in your class, or who you know well, who has special needs. Spend a few minutes recalling as much as you can about this pupil.

Now consider the situations in which it would be useful and appropriate to assess the pupil by different approaches – norm-referenced, criterion-referenced and ipsative-referenced assessment.

List all the different assessments that have been made of the pupil you are thinking about, and identify the reference basis for each (norm, criterion or ipsative). How closely does what actually happens match what you thought would be useful and appropriate?



HOW CAN WE ANALYSE ASSESSMENT DATA AT CLASSROOM OR SCHOOL LEVEL?

This section examines:

- what data is worth analysing and how it should be analysed
- what pre-analyses we can use to help
- whether different types of assessment data can be combined together.

What do you need to know?

What is worth knowing about a class of pupils is how well different groups of pupils are progressing. With this information, adjustments can be made to ensure that no group of learners is being disadvantaged. Such groups are commonly determined by their:

- age
- gender
- ethnicity
- prior attainment
- most recent target(s).

Sometimes other factors are also considered, such as their:

- term of birth
- sibling position or home-support capacity
- visual-spatial attainment (established through CAT tests for instance)
- preferred learning style, left brain/right brain orientation, etc
- perceptions of learning/teaching.

Ideas for analysis and use of data

1. CHOOSE YOUR DATA SET

Firstly, you need to look at the kind of assessment data that you would most like to use in the analysis – the major principle should be that it is of particular value to you. Do not simply opt to use data because it lends itself easily to analysis. Worse still, do not introduce a completely new set of assessments purely because they produce data which is easy to analyse. Stick to your principles, and opt for the data which produces the best balance for teaching and learning in terms of:

- validity and reliability
- formative and summative assessment.

2. FIND A COMMON DENOMINATOR IN THE DATA, TO ENABLE COMPARISONS TO BE MADE

If the assessments produce numbers, then this is easy. If the assessment you want to use is more like the example of highly valid reading conferencing (as illustrated in table 3 on page 19), then this becomes more difficult. One option is to look through all your notes of your assessments and ascribe a value to the judgements that you made from them. You may decide to give each assessment a value, as shown in the following example (this is an arbitrary illustration).

Very good progress, very positive attitude/motivation/engagement, high-attaining, clear about targets and likely to achieve all.	5 points
Good progress, very positive attitude/motivation/engagement, high/average attaining, likely to achieve all targets.	4 points
Good progress, positive attitude/motivation/engagement, average attaining, likely to achieve targets.	3 points
Reasonable progress, mixed attitude, variable engagement, likely to achieve most targets but needs motivation and more engagement.	2 points
Some progress but not enough. Insufficient engagement.	1 point
Little or no progress, mixed or poor attitude/motivation, little engagement or support from home/peers, likely not to meet most targets.	0 points

You can use a similar approach with other subjects and methods of assessment such as observations, where you use similar criteria in each (such as 'progress', 'motivation', 'engagement' and 'attitude' in the above illustration). *Any use of such systems necessitates regular moderation of judgements, in order to guarantee consistency and quality.* Does your school or department carry out moderations like these?

Progress measures

If the assessments are made against a scale of progression – even as broad as the national curriculum levels – then it is possible to use these to form a set of values. Most usefully, many systems now break the levels into thirds. These are usually expressed as 'c', 'b', and 'a', (for example '2c', '2b', and '2a'). This enables useful assessments and individual pupil and group targets to be set, and is also sufficiently discriminating to allow class teachers or departments to analyse the data.

You can then easily begin to make the following kinds of analysis. Packages such as Microsoft Excel can be used to create these charts within minutes.

TABLE 4

Average reading assessment score (Sept – Dec)



Table 4 shows the average scores generated by the categorised reading assessments. The teacher has broken them down by gender and by knowledge of English. From the table, one can clearly see that there is a gender issue that extends to learners with English as an Additional Language (EAL). Male EAL learners who are at a more advanced stage of English acquisition (stage 3) are scoring more highly than the average of all boys. The teacher has then looked at the average progress the pupils have made over the past 12 months. This is measured by their national curriculum Average Point Score (APS) which is calculated by assigning two points to each third of level progress – so 2c = 2, 2b = 4 and 2a = 6. (This is a method used by QCA in the Autumn Package, which features later on in this section.)

TABLE 5

NC progress (average NC points) reading

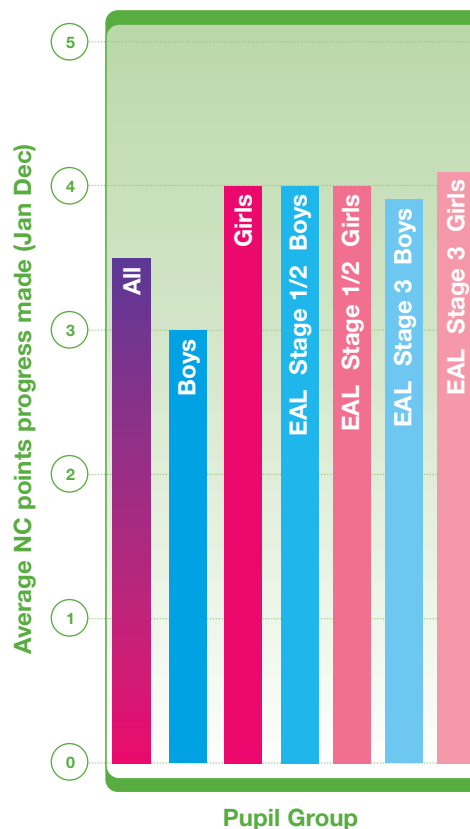
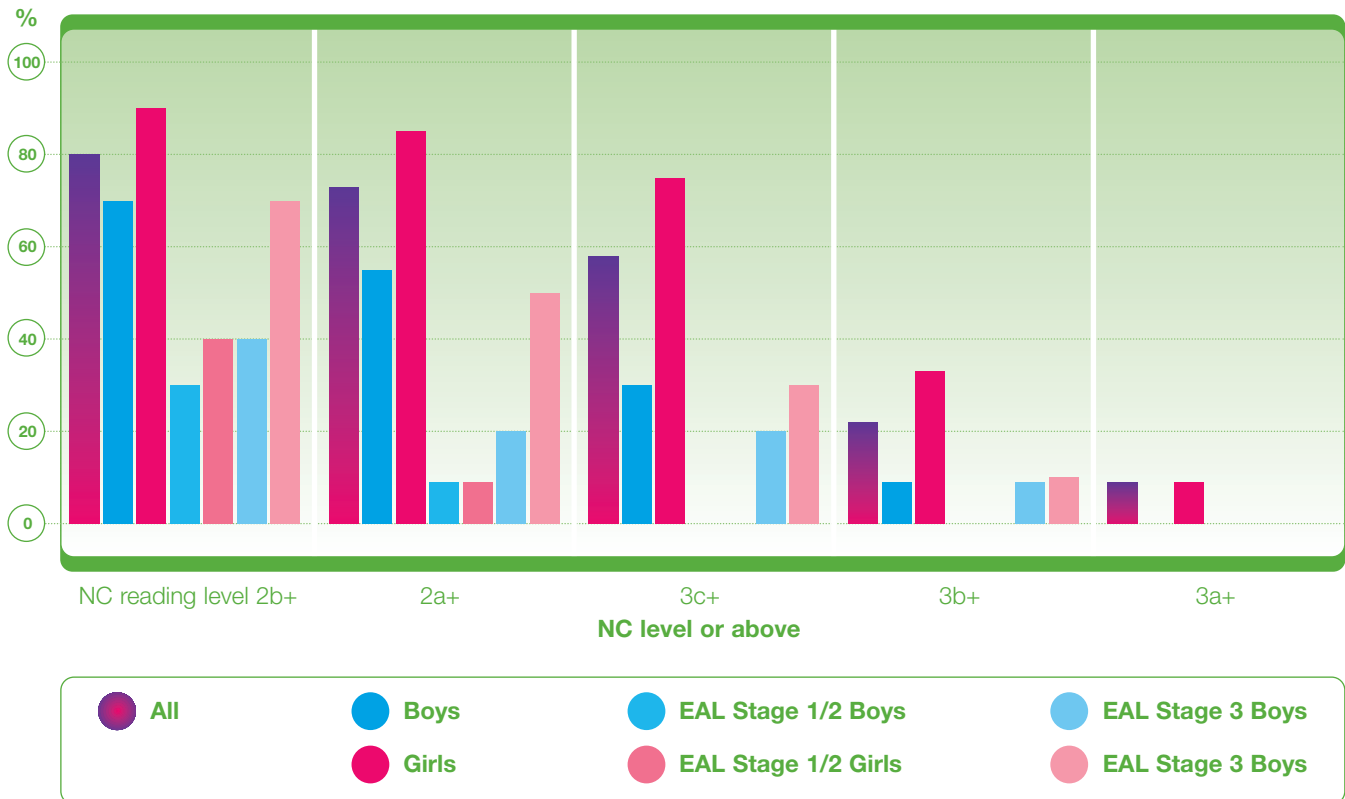


Table 5 shows the average progress for this Year 4 class is 3.5 points per year. This equates to a gain of well over one national curriculum level over two years. This is good progress in reading compared to the national averages. EAL pupils are generally making very good progress, despite their mixed reading assessments (see table 4).

TABLE 6

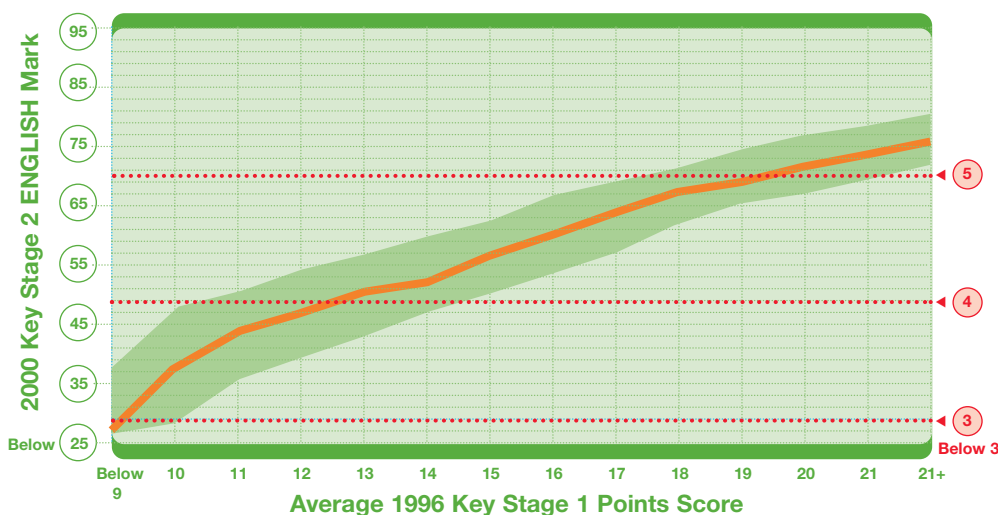
Attainment in reading



Finally, it is worth looking at the attainment levels of these groups as shown in table 6. The gender split detectable in both 'progress over time' and in 'reading assessments' is clear here. This also highlights the lower attainment levels which would normally be expected of EAL stage 1 and 2 learners in Reading English.

TABLE 7

Key stage 2 English value added line and progress charts



Autumn package for schools, DfES 2000

The Pupil Achievement tracker – formerly the Autumn Package for English schools (see above table) – breaks down end-of-key-stage attainment into each level by gender. It also provides average point score progress charts for the national samples of pupils, against which you can plot your own pupils to see how their progress over the key stage compares with national data. The space between the dotted lines represents the progress of 50 percent of Year 6 pupils nationally.

Similarly, for Years 3-5 it is possible to compare progress of pupils from the end of Year 2 using the QCA optional tests. The charts with the progression lines from the national data sample can be found on the QCA website (www.qca.org.uk/ca/tests/optional/tests_2000.asp).

ANALYSING TEST SCRIPTS

Analysing test scripts is a relatively simple task, which often bears fruit. Simply set out the question numbers along the top of your page and the pupil names down the side, and then allocate one mark for each item correctly scored by each pupil. If you group the pupil names by gender and/or in ascending age order, you may well see trends and patterns emerging. One infant teacher doing this discovered that, despite average results for her class in the key stage 1 tests, nearly all pupils – male and female – had failed to answer a level 1 item correctly.

The item was 'What is the difference between 8 and 12?' Pupils had not recognised the mathematical language that had been employed. Further analysis showed that this factor was affecting a number of responses to items, and so the teaching and modelling of mathematical language became a focus of INSET for the school the following year.

TABLE 8

Essex SATs analysis mathematics KS3 2001

Essex SATs analysis Maths KS 3 2001				Question values																			
				Male % errors																			
				Female % errors																			
				Total errors																			
All pupils				Question level																			
				Strand 2										Strand 1									
Forename	Surname	Gender	Class	7	1	5	6a	18	19	6	8	2	3b	6b	7	17	20	3	5	9	3a	3c	4
Harriet	Harman	F	9a	1	1		1	3	1	1	4			1	1		1	1	2	1	1	1	
Jill	Bertrum	F	9a		1	1	1	2			3	1		1	1	1	1	1		1		1	1
Julie	Harriet	F	9a				3															1	
Tom	Tipson	M	9a	1	1	1	1	2			4	1	1			1	1	1	2	1	1	1	1
Peter	Rabbit	M	9a																				
Roland	Rat	M	9a	1		1	1	2	1	1	3			1	1	1		1	2	1			1
Zippy	George	M	9a										1				1						1
Doug	Funny	M	9b					1															1
Filamena	Herbert	F	9b	1		1			1	2				1	1			1	2	1			1
Billy	Goat	F	9b					1											2				
Josie	Fill	F	9b			1			1	1	1			1	1	1		1	1	1			1
Susan	Peters	F	9b						1	1	2	1			1	1	1	1					1
Charlie	Farley	M	9b				1						1						1	1	1	1	1
Phil	McCavity	M	9b	1															1	1	1	1	1
Paula	Hills	F	9b				1	4		1	3	1		1	1	1			1				1

© Essex County Council Learning Services

The illustration above (table 8) is from a spreadsheet provided to schools by one LEA to help analyse key stage 3 test scripts and then interrogate the data by pupil groupings.

This spreadsheet allows subject leaders and teachers to analyse the items, and find out:

- in what areas of the tests pupils did well or less well at different levels
- how they did in different strands of the curriculum
- how different pupil groups (boys, girls, pupils from different classes) did in relation to different aspects of the test.

Using assessment data to set targets for pupil progress and attainment

Assessment data can be used to set targets for pupil progress and attainment by using the national curriculum-thirds of level intervals, in line with what is known about pupil progress through them. It is known that at key stages 1 and 2 in more effective classrooms, pupils progress at around three national curriculum points per year. At key stage 3, this is between two and three points. It is, therefore, possible to set such a rate as a reasonable target for each pupil. For those who do not relate to point scores, then the target could be expressed as two-thirds of a level per year.

These can be set out in your mark book at the start of the year next to the pupil's name. Individual learning targets can be agreed with the pupils, in order to set out their next achievement milestone. Table 9 below shows an example list of two-thirds of a level targets for a group of pupils over a year.

Cogent Software have calculated the following guide to targeted pupil progress for key stage 3. Similar guides are also available for key stage 1 and key stage 2. *The Times Educational Supplement* website, linked with Cogent, have a number of pages relating to the use of ICT and assessment. These pages can be accessed at www.tes.co.uk.

Please note that Cogent Software have used the term 'average' here to mean the expected/most common level attained by the cohort.

TABLE 9

Name	Attainment last July	Target level for next July
David	5a	6a
Mark	5b	6c
Sohaib	5b	6c
Leanne	6c	6a
Lisa	4a	5b

Some secondary schools now print such data on class lists, so all staff are therefore aware of prior pupil attainment and progress rates. Some also print a grading for consistency in 'effort'. 'Effort' is difficult to quantify or specify. In best practice, 'effort' is clearly defined to students. Examples of effort at different grades are provided and teacher assessments are monitored and moderated. An 'A' for effort is worth five points – an E, only one point. An average grading can be calculated for the year, and the current year's grades can be set against those from the previous year. Thus, it is easy to detect those pupils whose 'effort' seems to be slipping away. Early detection of a fall-off in motivation and engagement (effort) is important. Waiting until it shows in the attainment and progress assessments usually takes longer, and more damage will have been done.

In Autumn 2003 the DfES introduced the Pupil Achievement Tracker – an interactive pupil level analysis CD which allows a range of staff in schools to ask questions and study a variety of analyses of assessment data including 'value added' data and item analyses such as those featured in the Essex example on the previous page. At the time of writing it is too early to judge the impact of this development but an early view strongly suggests that it promises to improve greatly on the previous Autumn Package which could only use school level data and which relied upon schools to complete the pupil level analyses by hand.

The dangers with such a pupil level system will always lie in the quality and accuracy of the data that feeds the package – especially dealing with individual pupils and small cohorts. The greatest challenge, as with all data packages, is in ensuring that the staff who need to use the data in classrooms have access to it, and that they then understand the strengths and limitations of that data – and this is a significant training issue for LEAs and schools.

TABLE 10

Target indicator tables key stage 3- All National Curriculum subjects (except MFL)

	baseline ks2 results	end of year 7 NC interim target	end of year 8 NC interim target	ks3 NC target level	Other indicators Standard Scores
Special needs	-3	Targets for special needs pupils will be set by a different process			< 70
Well below average	3	3c	3b	4	70–85
Below average	3/4	3a	4b	4/5 (5 target)	86–93
Average	4	4a	5b	5/6 (6 target)	94–108
Above average	4/5	5c	6c	6	109–115
Well above average	5	5a	6a	7	116–130
Special needs – gifted children	5+	6c	7c	7+	> 130
end ks3					
4 and below (MFL 3 and below)	represent achievement below the nationally expected standard				
5 and 6 (MFL 4 and 5)	represent achievement at the nationally expected standard				
7 and above (MFL 6 and above)	represent achievement above the nationally expected standard				
sub-level c	indicates that a pupil is starting to work within the level				
sub-level b	indicates that a pupil is working well within the level				
sub-level a	indicates that a pupil understands most ideas within the level and is ready to move to the next				

At the end of each year, a level and sub-level are determined through standardised testing as determined in your school. Pupils who achieve a target higher than predicted will be placed in the band above for target setting the following year, so they will be working towards a higher end of key stage level.

Pupils who do not achieve the expected target need appropriate support to ensure that they remain on target the following year. **Only in very exceptional circumstances would a pupil be moved down a band.**

Pupil involvement in the target setting and self assessment process

The importance of pupil involvement in these processes should not be underestimated. Recent research has shown how:

- pupil involvement in self-assessment and target setting,
- pupil ownership of targets for which the learning strategies for achievement are understood, and
- a clear idea of what achievement or success will look like when it is attained,

are all vital in raising pupils' ability to 'learn to learn'. This is not the subject of this book, nor is there scope here to consider this in detail. However, if you are interested in a useful, practical summary of the research, please refer to *'Inside the Black Box'*,¹⁹ *'Working Inside the Black Box'*²⁰ and *'Beyond the Black Box'*²¹. There are also some informative websites; for more information, please refer to the further information section at the end of this book.

Comparing your pupils' assessment information with pupils in similar schools

The Autumn Package provides analyses of how well pupils have performed academically compared to those pupils with similar starting points (prior attainment). It groups schools according to the average point score of the cohort on entry to the current key stage, and then compares the progress pupils have made in one school with that of all schools in its prior-attainment-point-score group. The schools are then listed in the order they fall. The school that comes exactly a quarter of the way down the list is chosen to delineate the top quartile (or 25%). The school's score is the marker for this top quartile. All schools with scores above this school's score must have been above it in the list – so they are in the top quartile. Having established this, you

can then calculate your Ofsted PANDA (Performance and Assessment report for English schools) grade, based on where your score falls in comparison with these similar schools. In Wales, the National Benchmark Information sent to schools in November serves a similar function. The grades match percentile rankings as follows.

Top 5%	-A*
Top 25%	-A
Between 60% and 74%	-B
Between 40% and 59%	-C
Between 25% and 39%	-D
Between 5% and 24%	-E
Between 0% and 4%	-E*

The Autumn Package uses a similar process using the Free School Meal (FSM) entitlement to group schools and create similar gradings. This FSM comparison is likely to be phased out as prior attainment data becomes available – but despite its faults, the FSM comparison in many ways is still 'fairer' than simply comparing all schools against each other.

It is worth looking at Autumn Package/Ofsted grades, in particular if you are a head of department or a subject leader, as they are supplied to registered inspectors prior to inspection. Inspectors themselves will then make inferences about the quality of teaching and leadership which may be behind good or poor grades.

ACTIVITY 9

Check out your subject grading for each key stage and then make sure that you have a commentary explaining any variations (for example, the cohort had a low baseline) and a list of actions to improve attainment in any areas with low grades. See how this relates to your School Improvement Plan.

HOW DO WE MOVE FROM ANALYSIS TO ACTION?

No matter how wisely assessments have been made, and no matter how detailed and thoughtful the analysis of the data has been, learning and teaching will not improve after an assessment unless any actions indicated as necessary by that assessment are taken.

Any action needs to:

- address the issues raised by the analysis of assessment outcomes
- fit in with other developments in the school
- be well-planned
- be monitored and evaluated
- be worth the time and effort involved.

Addressing the issues

To a certain extent, the issues raised by analysis will depend upon the focus for that analysis in the first place. For example, EAL stage and gender differences in attainment and progress will not become apparent unless they were used as a basis for comparison.

It is important to identify the scale and scope of any issues which arise from the assessment. Is it a concern for a few pupils, quite large groups, whole classes, or across the school? Is it a question about one particular teacher, a year group, or a whole subject area? The answers to these questions will direct the most appropriate response to the analysis of the assessment outcomes. It may be that further investigation needs to be undertaken to answer some of these questions.

Concerns may have been raised through end-of-key-stage testing for example, and comparable information may not be immediately available about other year groups. Before deciding upon a major programme to address the apparently weak grasp of

'Materials and their properties', it would be worth remembering that this year group was the last one before the school switched to a new scheme of work where *'Materials and their properties'* has much greater prominence. In addition, due to long-term sickness, the class was taught by two supply teachers during the previous year, neither of whom were strong in this area of teaching. Looking at a sample of work from other year groups should then confirm whether or not the testing revealed a one-off difficulty that unfortunately was not rectified earlier, or a larger problem that does need to be tackled.

The opposite and equally inappropriate response to 'making a mountain out of a molehill' is to ignore unpleasant findings. There may be many reasons for doing this, but certainly no excuses.

Another inappropriate response is to regard some issues as those upon which the school can have no effect. The 'what can you expect with these pupils' syndrome has had a damaging effect on too many pupils, and whatever one teacher thinks, others have believed and proved that the same pupils can raise their attainment considerably. *'Success Against the Odds'* (National Commission on Education)²² and *'Success Against the Odds – Five Years On'* (National Commission on Education)²² both report on 11 schools which, although being located in disadvantaged areas throughout England, Scotland, Wales and Northern Ireland, have all been judged to be effective.

Fitting in with other developments

Hargreaves and Hopkins²⁴ talk about distinguishing between root and branch innovations, and forging links between priorities. Some things need to be in place before others can be really successful, and these sorts of considerations should determine the order in which certain developments need to happen. For example, it is pointless trying to encourage pupils to evaluate their own progress against learning intentions unless teachers are crystal clear about the learning intentions for each lesson and share them in a way that can be understood by the pupils.

An analysis of assessment data may point to actions which are related to other things already on the school improvement plan. For example, test item analysis showed that data handling was a particular area of weakness across a school. In that school's improvement plan, data handling had already been identified as an area for development, but this was not scheduled to take place until three terms later. In view of the effect that poor data handling skills were having, the school decided to bring forward its focus on data handling.

Planning well

As with any other developments, action arising from assessment should be planned well if it is to be effective. What this means in practice depends to a large extent on the scale of the action. A teacher deciding what to do in the classroom on the basis of having identified through her/his marking the same recurring problem, will plan differently from leaders tackling a school-wide issue. In each case though, objectives, success criteria, appropriate activities, timescale, monitoring and evaluation should all be considered. Whatever is planned should be manageable in itself, and also when taken together with other tasks.

Monitoring and evaluation

Monitoring and evaluation are essential for keeping track of how developments are progressing, and for judging their effectiveness. One of the dangers of any action is unintended negative outcomes, despite good intentions. It is only by monitoring that the actual effect, as opposed to the planned effect, can be judged. This is the case even if an issue has been successfully tackled in a particular way somewhere else: solutions may not work in another context.

Worth the time and effort?

There are many demands on teachers, and energy and effort need to be used wisely. Realising that it is very difficult to respond to all the issues raised by assessment may help us to think carefully about the resources expended on collecting assessment data and how much of it is actually analysed and used. Thoughtful responses to valid and reliable assessment information, arrived at through appropriate analysis, are what using assessment for wise decisions is all about.



7

FURTHER INFORMATION

The following websites, books and videos offer further information about assessment (as of September 2003).

Useful websites

Association of Assessment Inspectors and Advisers

www.aaia.org.uk/

Assessment Reform Group website

www.assessment-reform-group.org.uk

**DfES Pupil Achievement Tracker (from November 2003)
Autumn Package section of the Standards**

www.standards.dfes.gov.uk/performance/

EPPI-Centre

www.eppi.ioe.ac.uk

**GTC's research of the month on raising standards
through classroom assessment**

www.gtce.org.uk/research/standhome.asp

QCA Assessment for Learning

www.qca.org.uk/ca/5-14/afl/index.asp

The Times Educational Supplement websites/links

www.tes.co.uk/online/assessit/tesassessithome.htm

www.cogentcs.co.uk/performat.htm

**Centre for Research on Evaluation, Standards and
Student Testing (CRESST), at the University of
California in Los Angeles**

www.cse.ucla.edu

Useful books

'Assessing Children's Learning' Second Edition

2003, Drummond, Mary Jane, David Fulton,
ISBN 1-843120402

This book provides an important critical alternative to the current objective, mechanical approaches to assessment. Assessment is seen as a process during which teachers look at pupil's learning, strive to understand it and then make use of this knowledge in the interests of pupils.

'Using Assessment for School Improvement'

1998, James, Mary, School Management Series,
Heinemann, ISBN 0-435-80046-9

A framework for auditing and developing a whole school approach to assessment – useful to dip into and features many good practical suggestions. Examples from secondary schools are given.

'Enriching Feedback in the The Primary Classroom'

2003, Clarke, Shirley, Hodder and Stoughton,
ISBN 0-87258-6

Practical strategies that will engage with and support a pupil's learning, improve progress and raise confidence and self-esteem.

'Investigations. Targeted Learning'

2000, Goldsworthy, Anne, Watson Rod and Wood-
Robinson, Valerie, the Association for Science Education.

This book is based upon the finding of the ASE and Kings College Science Investigations in Schools Project, which was developed from the belief that scientific enquiry is at the core of school science. It focuses on the importance of formative assessment in scientific enquiry and the role of assessment in helping students in their ongoing learning.

‘Coordinating Assessment Practice Across the Primary School’ 1999

Harrison, Mike & Wintle, Mike, Falmer Press.
ISBN 0-7507-0698-8

This book is one of a series aimed at primary school coordinators. It provides a step-by-step introduction to the responsibilities of the assessment co-coordinator and includes practical guidance on many of the issues faced by teachers with this responsibility.

‘Investigating Formative Assessment. Teaching, Learning and Assessment in the Classroom’ 1998

Pryor, John, & Torrance, Harry Oxford University Press,
ISBN 0-335-19734-5

This book explores how the assessment of young children is carried out in the classroom and what the consequences of this are for their understanding of school and their learning. It is based on extensive video and audio recordings of assessment ‘incidents’ and interviews with teachers and pupils.

‘Statistics for School Managers’. 2000

Schagen, Ian Courseware Publications, Apple Barn Court,
Westley, Suffolk, ISBN 1-8987-3723-1)

A useful book for those keen to know more about statistics and use of attainment data.

‘Interpreting Pupil Performance Information: Knowing your PANDA from your PICSII!’ The National School Improvement Network Bulletin, No 11, Spring and Summer 2000

University of London Institute of Education

As the title indicates, this article attempts to clarify strategies for the analysis of performance data.

Videos

‘Assessment – making a difference’ and ‘Assessment for learning – a revolution in classroom practice’

Birmingham City Council Education Service

Useful training videos with supporting booklets. Strongly influenced by the work of Paul Black and Dylan William, the video puts a strong case for ‘assessment for learning’ with sections on the use of learning objectives, feedback to pupils and self/peer assessment. For more information: BASS Publications tel: 0121 303 8081.

References

- 1 Earl, L et al, 'Watching and Learning: OISE/UT Evaluation of the Implementation of the National Literacy and National Numeracy Strategies'. 2000. Ontario Institute for Studies in Education of the University of Toronto.
- 3 Assessment Reform Group (2002) 'Assessment for learning: ten principles'. Assessment Reform Group.
- 3 Stiggins, R. (1994) 'Student centred Classroom Assessment'. Toronto: Merrill.
- 4 Association of Inspectors and Advisors (AAIA) 2001 'Primary Assessment Practice: Evaluation and Development Materials'.
- 5 QCA (2000) 'English Tasks Teacher's handbook'.
- 6 DfEE (2000) 'Autumn Package 2000 GCSE/GNVQ'.
- 7 Messick, S. (1989) 'Meaning and values in test validation: the science and ethics of assessment'. *Educational Researcher* 18 (2) pp5-11.
- 8 Qualifications and Curriculum Authority, 'Key stage 2 tests in English, mathematics and science: level threshold tables and age standardised scores'. 2000.
- 9 Black, P. (1998) 'Testing: Friend or Foe? Theory and Practice of Assessment and Testing'. London: Falmer Press.
- 10 William, D. (2000) 'The meanings and consequences of educational assessments'. *Critical Quarterly*, 42 (1) pp105-127.
- 11 Kellner, P. (1997) 'Times Educational Supplement' 19 September 1997.
- 12 Harlen, W. (1994) 'Issues and Approaches to Quality Assurance and Quality Control in Assessment'. London: Paul Chapman.
- 13 SCAA (1996) 'Standards in Public Examinations 1975-1995'.
- 14 Hendry, J. (2001) 'Times Educational Supplement' 6 April 2001 .
- 15 Rose, J. (1999) 'Weighing the Baby: The Report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum tests in English and mathematics.' Report prepared for the Secretary of State.
- 16 DfEE (2001) 'Supporting the Target Setting Process'.
- 17 William, D. (2001) 'Level best? Levels of attainment in national curriculum assessment'. Association of Teachers and Lecturers.
- 18 Essex County Council (1999) 'Building Trust: Creating expectations. Initial Report of the Key Stage 2-3 Agreement and Year 6-8 Target Setting Project'.
- 19 Black, P., & William, D. (1988) 'Inside the Black Box: Raising Standards through classroom assessment'. London, Department of Professional Studies, King's College.
- 20 Black, P., Harrison, C., Lee, C., Marshall, B., and William, D. (2002) 'Working Inside the Black Box: Assessment for learning in the classroom'. London: Department of Professional Studies, King's College.
- 21 Assessment Reform Group (1999) 'Assessment for Learning: Beyond the Black Box'. University of Cambridge School of Education.
- 22 National Commission on Education (1996) 'Success against the odds: effective schools in disadvantaged areas'. London: Routledge.
- 23 National Commission on Education (2001) 'Success against the odds: Five years on'. London: Routledge.
- 24 Hargreaves, D. and Hopkins, D. (1991) 'The Empowered School'. London: Cassell.

The Association of Teachers and Lecturers exists to promote the cause of education in the UK and elsewhere, to protect and improve the status of teachers, lecturers, and non-teaching professionals directly involved in the delivery of education, and to further the legitimate and professional interests of all members.

For a free copy of the Association of Teachers and Lecturers' publications catalogue, please call the publications despatch line on 0845 4500 009.

© Association of Teachers and Lecturers 2003. Second edition. All rights reserved. Information in this book may be reproduced or quoted with proper acknowledgement to ATL.

To receive the text of this book in large font, please contact ATL on 020 7930 6441 or write to ATL, 7 Northumberland Street, London WC2N 5RD.

