The Role of Teachers in Teacher Assessment in England 1996–1998

Shirley Clarke
Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK

Caroline Gipps
Kingston University, London, UK

In this article the role of teachers in Teacher Assessment in England is explored. Teacher Assessments consist of professional judgements made at ages 7,11 and 14 against the criteria of an eight-level scale, *based on teachers' ongoing teaching rather than by testing*. These judgements are set alongside test results and have equal weight in reporting National assessment results, so they play a significant part in measuring standards. Through research projects carried out from 1996 to 1998, the authors analyse how teachers make Teacher Assessment judgements, how consistent these judgements are, and the extent to which consistency can be achieved. This is contextualised within the framework of international findings about related issues, where parallel systems exist.

International Issues

A number of countries have moved in recent years to implement national or state curriculum frameworks and assessment schemes which require teachers to report on student progress at designated times during primary and secondary schooling according to specific 'benchmarks' or 'standards'. These standards are represented as developmental steps, stages or levels, in some cases based on national curriculum frameworks. In these schemes teachers play a key role in collecting evidence of student achievement and interpreting this evidence in terms of the specified performance standards. In some cases, teacher assessments may be supplemented by external tests. These schemes represent a substantial change from past educational practice, replacing the previous psychometric paradigm of assessment (emphasising measurement, scaling and formal standardised tests), with the newer standards-based performance paradigm, emphasising authentic and contextualised assessment and involving teacher judgement and interpretations of standards (Maxwell & Gipps, 1996). More is demanded of teachers in standards-based performance assessment. However, little is known about the actual processes of assessment. We hope that this English study goes some way to enhancing understanding.

In Australia (McGaw, 1996) there is a similar curriculum structure to that in England and an eight level progressive system of performance outcomes.

> Curriculum development has long involved specification of scope and content and some form of declaration of the learning objectives held for students. What is new, in places like England and Australia at least, is an attempt to develop a more explicit standards perspective in curriculum by specifying student learning outcomes in developmental sequences.

0950-0790/00/01 0038-15 $10.00/0

> These sequences become specifications of standards when expectations of rates of student development are imposed on them as well. This can involve mapping of grade levels onto the sequences to indicate what some proportion of students is expected to have achieved by the end of each grade level or set of grade levels' (McGaw, 1996: 3)

As in England, consistency across teacher judgements is a major issue:

> The outcome statements offer teachers a constant language for thinking about student learning and for discussing it with students and parents. It gives teachers the chance to use consistent criteria as a reference for student achievement.
>
> The question is, can they use the criteria consistently with respect to other teachers or some group of experts, defined as 'experts' because they can make consistent, independent judgements of student performance against outcome scales? (McGaw, 1996: 13)

What McGaw's research has found is that teachers at different grade levels interpret the outcome statements in different ways: they make finer judgements in the range of outcomes in which their own students predominantly operate. In England, too, award of the same level at different key stages is an issue, as teachers of older ages may not accept that pupils from two or more years below can be operating at the level appropriate for the older key stages. In our research[1] we found that the majority of teachers were in favour of the eight-level scale (graded level descriptions), but felt that the levels were too broad. (The eight levels cover nine years of school in England and Wales; ten years of school in Australia).

Many of our teachers said that the eight-level scale was *not* being used as a continuum because of the difficulties of comparing pupils with the same level across different key stages and because of the hiatus which occurs when pupils transfer from one school to another: pupils may be 'stuck' for a number of years because the levels in the previous school are not seen as realistic by the receiving school.

In New Zealand, with a similar assessment and curriculum structure, research has addressed 'over-assessment' (Keown, 1996) in relation to assessment of pupils against unit standards. The standards have a pass/fail structure and the research was concerned to prevent secondary teachers from doing too much formal assessment and to encourage them to build on 'naturally occurring' evidence instead: i.e. a concept of sufficiency:

> Quality sufficiency decision making can be defined as a process of collecting the quantity and quality of evidence required to convince an assessor that a candidate is or is not competent in relation to the function defined by the element without over assessing or under assessing. (Keown, 1996: 3)

In the Keown study, teachers' main concerns were: the quality of response required to award credit; quantity of evidence required to award credit; the problem of time for assessment and reassessment; how to collect naturally occurring evidence fairly; authenticity and how to cope with group work and homework; how to track, record and feedback evidence and performance quickly enough for students to be able to benefit and present better at a re-sit; and how to get consis-

tency between teachers and between schools. The author concluded that there is a need to institute a programme of training to assist teachers to broaden their repertoire of assessment strategies so that they can gather valid, naturally occurring evidence to supplement their formal assessment activity evidence, thus reducing the amount of formal assessment required.

However, the use of naturally occurring evidence is not so simple, as the findings from our study indicate. If the use of assessment results is high stakes then reliability and consistency issues come in to full play. *Consistency of standards* relates to ensuring that different teachers interpret the assessment criteria in the same way, whether using naturally occurring evidence or setting tests. However, where tests are used it may be necessary to ensure *consistency of approach*: the assessment task or activity which is used and the way in which such tasks are presented to the pupil, or indeed contextualised, can effect performance quite markedly. To ensure consistency of approach, therefore, we need to ensure that teachers understand fully the constructs which they are assessing (and therefore what sort of tasks to set); how to get at the pupil's knowledge and understanding (and therefore what sort of questions to ask); and how to elicit the pupil's best performance (the physical, social and intellectual context in which the assessment takes place). This, of course, is a tall order.

Group moderation is a key element of teacher assessment, not only in terms of improving inter-marker reliability, but also to support the *process* of assessment. In group moderation samples of work are discussed by groups of teachers against the assessment criteria (i.e. level descriptions). This is sometimes repeated at a district or county level, with the results brought back to each school to achieve a broader consensus. This procedure is used in Queensland where assessments are made against a five-level scale and no external examinations exist. The enhanced validity offered by teacher assessments is gained at a cost to consistency and comparability (Mislevy, 1992), which can never be as consistent as the highly standardised procedures involved in testing.

If we wish to be able to 'warrant assessment-based conclusions' without resorting to highly standardised procedures with all that this implies for poor validity, then we must ensure that teachers have common understandings of the criterion performance and the circumstances and contexts which elicit best performance: this can be developed through group moderation.

The disadvantage of group moderation is that it is time consuming and costly and this may then be seen to add to unmanageability in an assessment programme. Its great advantage, on the other hand, lies in its effect on teachers' practice (Linn, 1993; Radnor & Shaw, 1994). It has been found that where teachers meet to discuss performance standards, or criteria, the moderation process becomes a process of teacher development with wash-back on teaching. It seems that coming together to discuss performance or scoring is less personally and professionally threatening than discussing, for example, pedagogy. But discussion of assessment does not end there: issues of production of work follow on and this broadens the scope of discussion and impacts on teaching (Gipps, 1994: 80).

The National Curriculum in England

The Education Reform Act of 1988 introduced, for the first time in recent

history, a national curriculum for children aged 5–16 together with a national assessment programme for pupils at ages 7,11,14 and 16.

The national curriculum was designed to ensure that all pupils of compulsory school age would follow the same course with English, mathematics and science forming the core, and history, geography, technology, information technology, a modern foreign language, art, music, design technology and physical education – the foundation subjects – forming an extended core. For each subject the curriculum is enshrined in law: statutory orders describe the matters, skills and processes to be taught as 'programmes of study' and the knowledge, skills and understanding as 'attainment targets' which pupils are expected to have reached at certain stages of schooling. The stages are defined as Key Stage One (age 5–7), Two (age 7–11), Three (age 11–14) and Four (age 14–16).

The national assessment programme was set up as a crucial accompaniment to the National Curriculum in 1989 in England and Wales, for it was through the assessment programme that standards were to be measured and eventually raised; the first stage of the development of the national curriculum was the setting up of the Task Group on Assessment and Testing (TGAT). The report of this group (DES, 1988) put forward a blueprint for the structure of the curriculum to which all subjects had to adhere. Subjects are divided into a number of components called attainment targets which are articulated at a series of progressive levels. The series of levels is designed to enable progression: most pupils of 7+ would be at level two in the system while most pupils of 11+ would be level four and so on. The attainment targets were described at each of the levels by a series of criteria or statements of attainment which formed the basic structure of a criterion referenced assessment system.

There are two main statutory assessment methods: external tests or assessment tasks and teachers' own informal assessments of pupils' attainment: Teacher Assessment (TA). For Teacher Assessment, teachers make an assessment of each pupil's level of attainment on the scale of levels in relation to the attainment targets of the core subjects. Teachers make these assessments in any way they wish, but observation, regular informal assessment and keeping examples of work are all encouraged. Statutory moderation of Teacher Assessment by LEA appointed auditors was set up in the first year of statutory testing at Key Stage One to ensure that TA judgements were 'approved'. This was subsequently removed because of the 'workload' problems which led to a national teacher boycott of all aspects of national testing in 1993.

Because of the reliance on Teacher Assessment, the TGAT report suggested a complex process of group moderation through which teachers' assessments could be brought into line around a common standard. Group moderation, commonly referred to as 'agreement trialling' in England, describes the moderation of teachers' assessments by the common or consensus judgement of a group or panel of teachers and/or experts or moderators (SSABSA, 1988). Group moderation relies on teachers' professional judgement and is essentially concerned with quality assurance and the professional development of teachers. In group moderation examples of work are discussed by groups of teachers; the purpose is to arrive at shared understandings of the criteria in operation and thus both the processes and the products of assessment are considered. The process can be extended to groups of schools within a district or county: samples of

graded work can be brought by one or two teachers from each school to be discussed at the district/county level. This will reveal any discrepancies between schools and the same process of discussion and comparison could lead to some assessments being changed in the same way as at the local level meeting. Teachers then take this information back to their own schools and discuss it in order to achieve a broader consensus.

Since the early 1990s, 'league tables' have been published in national newspapers ranking schools by their test results of 11- and 14-year-olds. This is a controversial practice in the UK, because, although value added results (where scores are altered to take account of aspects such as socioeconomic factors) are also published, schools are judged by various agencies, including parents, by their raw score results.

Whether to combine TA and test results, and how both are reported (both to parents and in public league tables) has been a contentious area: the rule at first was that where an attainment target was assessed by both TA and test and the results differed, the test result was to be 'preferred'. Currently the TA and test results are reported separately and, in theory, have equal weighting, although the league tables are based on the test results alone.

Boycott and review

In 1993, all three key stage tests were boycotted by teacher unions because of the perceived extra workload of administering and marking the tests and deciding TA levels. As a result, the Government set up a committee under the chairmanship of Sir Ron Dearing to review the entire national curriculum and assessment programme with the express aim of finding ways to simplify the testing programme.

The major outcomes of the Dearing Review were:

- a simplification of the curriculum;
- the suspension of league tables of schools' performances at ages 7 and 14;
- the reporting of TA alongside test results and giving both equal status (rather than subsuming the teachers' assessments);
- the removal of statutory moderation of TA levels by Local Education Authorities;
- a shift away from multiple statements of attainment to broad level descriptions.

Level descriptions/standards

In order to simplify the criterion-referenced basis and to reduce the task of tracking pupil attainment, almost 1000 statements of attainment have been reduced to 200 level descriptions. An example is given below:

> Attainment Target 2: Number and Algebra
> *Level 2*
>
> Pupils count sets of objects reliably, and use mental recall of addition and subtraction facts to 10. They have begun to understand the place value of each digit in a number and use this to order numbers up to 100. They choose the appropriate operation when solving addition and subtraction prob-

lems. They identify and use halves and quarters, such as half of a rectangle or a quarter of eight objects. They recognise sequences of numbers, including odd and even numbers.

These level descriptions are similar to the standards being used in parts of Australia and those being developed in the USA.

One anxiety about the level descriptions is that they are too global to be used as assessment criteria. If teachers are to use them for assessment purposes in anything more than a rough and intuitive way, they may break them down, an undesirable approach, as it could take teachers back to the unmanageable pre-Dearing 'tracking' regime; exemplars are also necessary in order to help classroom teachers make assessment against descriptions.

Despite official endorsement of the role of teachers in making assessment within the national assessment programme there has been limited support for teachers to undertake this. Some 'non-statutory' assessment material has been provided to schools and this has proved popular. Exemplification materials, to support group and individual national judgements about levels of performance have also been produced (e.g. SCAA, 1995).

The Research

Against this background we undertook three research projects for SCAA (The Schools' Curriculum and Assessment Authority), now known as QCA (The Qualifications and Curriculum Authority). The first, in 1996, was to monitor the consistency of Teacher Assessment in England across Key Stages 1, 2 and 3 and the extent of use of the centrally provided materials (Gipps & Clarke, 1996). That project was followed, in 1997 and again in 1998, by others to evaluate national assessment at Key Stage 1, including both tests and Teacher Assessment (Clarke & Gipps, 1997, 1998).

For the consistency project (1996) data from a total of 288 questionnaires from Year 2 (age 7) teachers, Year 6 (age 11) teachers, Assessment Coordinators and Heads of English, mathematics and science departments in secondary schools were analysed. Twenty-four schools were visited as case studies and a total of 77 teachers were interviewed.

In the evaluation project of 1997, a total of 212 questionnaires were returned from Year 2 teachers and 216 from headteachers; 20 schools were visited as case studies; the headteacher and the Year 2 teacher were interviewed in each school and at least one test was observed. The 1998 evaluation project used the same methodology. A total of 178 questionnaires were returned from Year 2 teachers and 174 from headteachers.

In all three projects we investigated how teachers make Teacher Assessment judgements and how consistent these judgements might be. That data is brought together in this article. Year 2 and Year 6 teachers were asked about their own practice. Heads of subject departments in secondary schools were asked about practice amongst specialist teachers in their departments.

There are three dimensions to making Teacher Assessment judgements in English schools:

(1) ongoing, day to day assessment judgements;

(2) end of Key Stage (ages 7, 11 and 14) summative level judgements made for each child for the core subjects;
(3) whole school or department standardisation meetings to ensure consistency of interpretation of level judgements.

We will present our data in relation to each of the three dimensions.

Ongoing, day to day assessment

In the 1996 consistency study, teachers were asked to describe the elements of ongoing teacher assessment which allowed teachers to make an overall level judgement at the end of the year for internal purposes or at the end of the Key Stage for reporting to parents. A list of possible strategies was provided for teachers to tick as well as space to describe other strategies. Table 1 shows that primary schools and English departments of secondary schools have many aspects of their ongoing assessment practice in common. In general, it seems that mathematics and science departments in secondary schools adopt rather formal approaches to ongoing assessment (e.g. end of module tests, regular classroom tests), whereas English departments and primary teachers tend to use more informal, formative methods (e.g. pupil self-assessment, regular notetaking, use of pupil portfolios).

Table 1 How teachers make ongoing assessment judgements

Elements of ongoing teacher assessment	*Year 2 teacher*	*Year 6 teacher*	*Head of English dept*	*Head of maths dept*	*Head of science dept*
Ongoing marking	56 93.3%	44 95.7%	33 97.1%	29 93.5%	21 84%
Regular informal assessments as part of the teaching plan	56 93.3%	42 91.3%	31 91.2%	26 83.9%	18 72%
Regular classroom tests	32 53.3%	37 80.4%	18 52.9%	26 83.9%	20 80%
Tracking significant achievement via a pupil portfolio	42 70%	30 65.2%	30 88.2%	15 48.4%	8 32%
Aspects of planning systems	44 73.3%	35 76.1%	19 55.9%	13 41.9%	10 40%
Involving pupils in self evaluation	29 48.3%	31 67.4%	29 85.3%	13 41.9%	13 52%
Regular collections of annotated samples	35 58.3%	23 50%	15 44.1%	10 32.3%	8 32%
Regular note-taking from structured or unstructured observations of practical and/or oral work	32 53.3%	28 60.9%	21 61.8%	5 16.1%	9 36%
Check lists based on level descriptions	30 50%	25 54.3%	14 41.2%	12 38.7%	7 28%
End of module tests with agreed criteria for the level to be awarded	15 25%	17 37%	9 26.5%	14 45.2%	20 80%

End of Key Stage TA summative level judgements

Evidence used to determine Teacher Assessment levels

The 1996 consistency study revealed that a variety of sources are used by teachers when deciding levels, as shown in Table 2. Most teachers, at this stage, said that the statutory test levels had no influence over Teacher Assessment levels (results are known before Teacher Assessment levels have to be completed).

Table 2 The evidence used in 1996 to determine teacher assessment levels

	Primary		*Secondary*		
Information used	***Year 2 teacher***	***Year 6 teacher***	***Head of English dept***	***Head of maths dept***	***Head of science dept***
General written work	58 96.7%	45 97.8%	33 97.1%	27 87.1%	22 88%
Set classroom tests or assessment activities	59 98.3%	40 87%	29 85.3%	30 96.8%	24 96%
Observations	59 98.3%	44 95.7%	29 85.3%	16 51.6%	20 80%
Dialogue with the pupil	54 90%	40 87%	25 73.5%	17 54.8%	14 56%
The pupil portfolio	32 53.3%	26 56.5%	28 82.4%	13 41.9%	10 40%
Memory	40 66.7%	25 54.3%	12 35.3%	14 45.2%	10 40%
Homework	8 13.3%	10 21.7%	22 64.7%	21 67.7%	18 72%

Most teachers used general written work and regular classroom tests or assessment activities when deciding levels. Most teachers also used observations of pupils as a source of information, with the exception of heads of mathematics departments. Primary teachers and heads of English departments were more likely to consider dialogue with the pupil as a source of information, which may reflect the lack of opportunity for dialogue in mathematics and science departments. The pupil portfolio was used by around half of primary schools and particularly in English departments. Memory was more likely to be used by primary teachers, which may reflect the fact that primary teachers have the same class all year, so have a great deal of knowledge about their pupils. Understandably, homework was used much more by secondary teachers as a source of information.

The 1997 Key Stage 1 evaluation study showed that the most common type of evidence used in Year 2 classes was teachers' records and children's work. 'Teachers' records' is a term likely to have many definitions so, in order to be clearer about the type of evidence used, this question was left open in 1998, but

Table 3 The evidence used by Year 2 teachers in 1998 to decide teacher assessment levels

Evidence used	*Number of teachers*
Ongoing and termly or half termly tests (mainly mathematics and science), either in-house or commercial	90
Pupil portfolios of annotated work, levelled in many cases	79
Jottings of ongoing assessments: achievements made, help needed etc. (weekly/daily)	74
Discussion/moderation with colleagues in school	73
Marking comments	54
Children's work	43
Level descriptions (some used as checklists)	35
Professional judgements based on knowledge of the child	28
Observational notes (mainly science)	19
Discussion with the children	14
School portfolios	12

with suggestions made of the level of detail required in teachers' responses. Table 3 shows the most common strategies given.

In contrast to 1997, ongoing tests seem to be a feature in many infant classrooms by 1998, not only as a means of preparing children for statutory tests, but also to inform Teacher Assessment summative judgements.

Of the process Attainment Targets for mathematics, science and speaking and listening (known as the AT1s), 'Speaking and Listening' for English and problem solving and investigation for mathematics and science have been found in various SCAA evaluations to be the most difficult aspects of the curriculum both to teach and to assess. Indeed, the next section of this paper describes how mathematics and science departments in secondary schools make AT1 the focus of all their standardisation meetings. The 1997 questionnaire asked separately about deciding TA levels for AT1; the data showed that, even as early as Year 2, teachers use specific assessment tasks in order to gather evidence for Attainment Target 1 for mathematics and science. For speaking and listening, however, the strategies are more widespread, including use of set tasks and discussion with colleagues, while memory is the most popular option.

Year 2 teachers were asked, in both 1997 and 1998, via the questionnaire, to say whether they had used any of the SCAA/QCA test criteria to help them decide TA levels: in 1997 70.5% of teachers said they had used them (70.8% in 1998). The task and test criteria (which are not the same as the level descriptions) are written in order to judge one piece of work, rather than for overall performance in an Attainment Target across a range of contexts. The use of tasks and test criteria for TA is a symptom of the lack of clarity in the level descriptions and of the 'best fit' approach. The writing task performance descriptions were most used, the main reason given that the criteria are much clearer than the Level Descriptions and because, as one teacher put it, 'They get into your consciousness'.

The best fit approach

End of Key Stage TA level judgements are supposed to be based on the level descriptions from the 8-level scale of the Attainment Targets for the various subjects. The statutory advice for determining a level is to apply a 'best fit' notion, which

> is based on knowledge of how the pupil performs across a range of contexts, takes into accounts strengths and weaknesses of the pupils' performance and is checked against adjacent level descriptions to ensure that the level awarded is the closest match to the child's performance in each attainment target. (QCA/DfEE, 1998)

The 1996 findings showed that most teachers did not think that the 'best fit' approach worked very well, because it was difficult to make decisions about pupils who appeared to fall between two levels and the notion of 'best fit' was too vague. However, having just been released from the previous system of counting the number of Statements of Attainment a pupil had achieved in order to determine a level, teachers said that they found the approach much more manageable, so did not want it to be changed.

The 1997 evaluation found that, although Year 2 teachers still considered the approach manageable and did not want to return to the previous unmanageable system, another year's experience of making 'best fit' judgements had made them feel that it was not a good means of representing children's achievements. Questionnaire comments revealed that teachers felt that the approach was too open to different interpretations across schools. These findings were mirrored in 1998, although fewer teachers said that the approach was manageable.

It also emerged that teachers were dissatisfied with the eight-level scale (graded level descriptions for each Attainment Target, intended to span the age ranges 6–14) in providing a continuum of performance. Although this is part of a complex picture, which includes the influence of league tables, it seems that secondary teachers tend not to believe the levels sent up to them by primary teachers, due to perceived over generosity and lack of subject knowledge. Teachers also found it difficult to consider the levels without taking account of the age of the child, and the accompanying Programmes of Study for the age group. As one teacher put it:

> How can you relate a Level 3 five- or six-year-old to a Level 3 fifteen-year-old? The disputes come from the structure itself: it means something different for different ages.

How teachers interpret 'best fit'

Bearing in mind that 1996 was the first year of determining TA levels in this way, we attempted to find out exactly how teachers were interpreting and applying the concept of 'best fit'. A number of statements were given for teachers to tick if they agreed. As illustrated in Table 4, most teachers said that they made 'general best fit judgements'. Primary teachers and heads of English departments were more likely to use best fit judgements in relation to children's portfolios. (This links with the earlier findings about formative assessment strategies). Approximately half of each group of teachers said that they identified key aspects of level descriptions (individuals must be able to do x, y and z in order to

Table 4 How teachers make 'best fit' judgements

	Primary		*Secondary*		
	Year 2 teachers (60)	***Year 6 teachers* (46)**	***Head of English (34)***	***Head of maths (31)***	***Head of science (25)***
By making general 'best fit' judgements	43 71.7%	35 76.1%	18 52.9%	17 54.8%	18 72%
By using 'best fit' judgements in relation to children's portfolios	35 58.3%	22 47.8%	23 67.6	10 32.3%	5 20%
By splitting the level descriptions (e.g. by creating separate statements and counting half or more as attaining a level)	12 20%	8 17.4%	4 11.8%	5 16.1%	3 12%
By identifying key aspects of level descriptions	31 51.7%	23 50%	14 41.2%	8 25.8%	13 52%

reach this level) in order to determine 'best fit', with the exception of heads of mathematics departments, where only 26% of teachers said that they did this.

Interviewed teachers were also asked how they had used level descriptions to arrive at a level. Responses were very varied, but the overall picture was of secondary teachers averaging the set of levels which pupils had by the end of the year and primary teachers using a best fit judgement. We felt it would be interesting to pursue this further, to try to establish exactly how primary teachers define 'a general best fit judgement'. So more options were given to Year 2 teachers in both 1997 and 1998. The findings for 1997 are shown in Table 5.

Table 5 How Y2 teachers interpret 'best fit' (1997) (N = 212)

'Best fit' interpreted as	***Yes***	***No***
The level description which overall describes the child's attainment better than the one above or below	71.7% (152)	28.3% (60)
Must achieve 75% or more of the statements in the level description	44.3% (94)	55.7% (118)
Must achieve important aspects of a level description	25.9% (55)	74.1% (157)
Intuition	17% (36)	83% (176)
Must achieve almost 100% or 100% of the statements in the level description achieved	15.1% (32)	84.9% (180)
Must achieve 50% or more of the statements in the level description	1.9 % (4)	98.1% (208)
Other	1.4% (3)	98.6% (209)

The table shows that the most common interpretation of 'best fit' is to decide which level describes the attainment of the child more appropriately than adjacent levels. This statement was put in specifically at the request of SCAA officers and against our advice, since it does not tell us how the teacher makes the decision as to what is 'appropriate'. In order to decide that one level is more appropriate than another, some judgement has to be made, such as deciding key indicators or counting statements attained, or alternatively intuition. By 1998 we asked teachers to qualify this option in the questionnaire, resulting in 79.3% of teachers saying that the child had to attain 75% or more of the statements at a level. The data from both years shows, however, that most teachers use more than one strategy to decide the level, the most common being using 75% of the statements and also identifying key aspects of a level. Teachers interviewed in 1997 showed an alarming lack of consistency on this question while broadly supporting the questionnaire findings. By 1998, opinions were only slightly less varied. Of the 20 teachers interviewed in 1998, 13 said that they looked for 75% or more statements, 7 said they looked for key indicators, while 2 said they expected 100% of the statements to be attained.

Whole school or department moderation meetings

This topic was pursued through interviews in the 24 case study schools in 1996. The interviews probed issues of consistency of level judgements, school organisation and effectiveness of moderation meetings, the usefulness of the government's guidance materials and how teachers arrived at 'best fit' judgements. Moderation meetings, in which samples of pupils' work are discussed with reference to given level descriptions to ensure a common interpretation of standard (also known as agreement trialling in the UK) are the main vehicle for enhancing consistency. However, the data revealed quite different approaches to these meetings between primary and secondary schools.

Secondary schools appeared to use moderation meetings in order to check on marks awarded for school-based tasks or tests, which were then used to determine the final level for a pupil. English departments tended to use marked samples of pupils' writing which were analysed against the level descriptions at the meeting. Mathematics and science departments set Attainment Target 1 tasks (investigations or problems) from three to six times a year, then set up meetings specifically to check the grades awarded against levels. (It appeared that these were the only times when pupils encountered Attainment Target 1 work, even though the National Curriculum demands that investigative work be an ongoing feature of lessons.)

By contrast, primary schools used a range of pupils' ongoing work as the focus for moderation meetings, analysing the work and *deciding a school interpretation* for the definition of a level. The work was then often put into a school portfolio, to be used for reference when deciding levels. Arriving at a standardised interpretation of levels was used by primary teachers for more than deciding final TA levels: having a clear view of the progression from level to level for each attainment target helped teachers to plan work which was appropriate for their age group. All the teachers, in that study, found the meetings, whether whole department or school, or of a small group, very useful and effective.

The 1997 evaluation showed that primary schools were having fewer modera-

tion meetings of the kind described in the 1996 data, but were tending to focus more on standardisation meetings specifically to discuss borderline cases for the writing task in the statutory tests: at Key Stage 1 the teacher has to decide the test result, as these papers are not externally marked. Thus it appears that the needs of the test have changed the focus of meetings, although Year 2 teachers' increasing familiarity with the level descriptions and use of 'best fit' has probably led to a perception that there is less need for standardisation meetings in which level interpretations are discussed.

Discussion

A key element of national assessment policy in England is the involvement of teachers in assessing pupils' performance. This is important for two reasons: to give teachers a stake in the assessment process and to allow assessment of a broad range of skills and processes in order to maintain a broad curriculum. Only certain attainment targets are tested by national tests, and, inevitably, those which are more difficult to test (speaking and listening, process elements, investigation and problem solving) are left for Teacher Assessment only. Teachers' comments about Teacher Assessment made it clear that they find the mechanisms (e.g. record keeping, moderation meetings) time consuming. However, they think it is an essential process which has a direct impact on pupils' learning and their teaching: teachers indicated that, regardless of the workload problem, they wish to continue with Teacher Assessment in all its forms.

This research highlights the complexity of making what are essentially reporting judgements against broad descriptions of performance or performance standards.

The notion of 'best fit' is a consciously loose one. Because of this, teachers are taking a variety of approaches to making Teacher Assessment judgements. Some teachers will make quantitative judgements (to attain a level individuals must meet all elements of a level description, 50%, or some other proportion); some will take a hurdle approach (individuals must do x, y and z in order to reach Level 5); others will take an intuitive approach (this one feels like a good Level 4). Although not addressed in this study we know that some teachers will make ranking judgements (this is a clear Level 7, and this is a clear Level 6; less clear performances are then slotted in, in relation to these fixed points). Because of the lack of clarity of 'best fit', the differences in interpretation mean that, at times, there will have been a difference of one level awarded to pupils, and this is not acceptable in a 'high stakes' programme.

Our findings point to the fact that moderation meetings or agreement trialling, especially cross-school and cross-phase, are particularly important, as moderation is a crucial process to achieve consistency. Teachers clearly value Teacher Assessment and see its importance in maintaining a broad taught curriculum. They see moderation meetings as valuable (despite the time issues); and primary teachers, especially, would value cross-school agreement trialling. Other research has clearly shown that consistency in teacher assessment can best be achieved by use of exemplification materials *and* some form of group moderation (see Harlen, 1994).

The differences in moderation practice and use of exemplification materials

across primary and secondary schools are in large part due to the experience of the General Certificate of Secondary Education (GCSE) at secondary level which has involved secondary teachers in assessing and moderating pupils' work for many years. The examination boards which produce and mark these exams offer materials and procedures which are widely used. However, if secondary mathematics and science teachers could be encouraged to use more informal, formative assessment methods, rather than relying on tests, more valuable feedback could be provided to pupils. The most accessible of the strategies used by secondary schools was involving pupils in self assessment, via target setting and the sharing of learning intentions. This resonates with good practice in learning.

The development of assessment skills among English primary teachers shows how it is possible to give teachers a central role in any assessment programme. In 1990 much of their knowledge about assessment was rudimentary and their practice intuitive; tremendous development has taken place (Gipps *et al*., 1995; Brown *et al*., 1997). But such development takes time, professional development and support material.

It is possible, and we would argue desirable, to give teachers a role in an assessment programme but the judgements made and the underlying requirements are complex. We hope this paper has illuminated some of the issues around evidence, judgement and consistency in such teacher assessment practice.

Correspondence

Any correspondence should be directed to Shirley Clarke, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK (sclarke@forestview0.demon.co.uk).

Note

1. Monitoring consistency in teacher assessment and the impact of SCAA's (Schools' Curriculum and Assessment Authority) guidance materials at Key Stages 1, 2 and 3 March 1997 – Report to SCAA 1997 (QCA Publications). Evaluation of Key Stage 1 Statutory Assessment (England) 1997 – Report to SCAA 1997. Evaluation of Key Stage 1 Statutory Assessment (England) 1998 – Report to QCA(Qualification and Assessment Authority) 1998 Website: www.qca.org.uk.

References

Brown, M., McCallum, B., Taggart, B. and Gipps, C. (1997) The validity of national testing at age 11: The teacher's view. *Assessment in Education* 4 (2), 271–293.

Clarke, S. and Gipps, C. (1997) *Evaluation of Key Stage 1 Statutory Assessment (England) 1997*. London: SCAA.

Clarke, S. and Gipps, C. (1998) *Evaluation of Key Stage 1 Statutory Assessment (England) 1998*. London: QCA Website.

DES (1988) *National Curriculum: Task Group on Assessment and Testing: A Report*. DES/Welsh Office.

Gipps, C. (1994) Quality in teacher assessment. In W. Harlen (ed.) *Enhancing Quality in Assessment* (pp. 71–86). London: Paul Chapman.

Gipps, C., Brown, M., McCallum, B. and McAlister, S. (1995) *Intuition of Evidence? Teachers and National Assessment of Seven Year Olds*. Milton Keynes: Open University Press.

Gipps, C. and Clarke, S. (1996) *Monitoring Consistency in Teacher Assessment and the Impact of SCAA's Guidance Materials at Key Stages 1, 2 and 3*. London: SCAA.

Harlen, W. (ed.) (1994) *Enhancing Quality in Assessment*. BERA Policy Task Group on Assessment, Paul Chapman Publishers.
Keown, P. (1996) *Walking the Sufficiency Ridge*. Final Report of the Sufficiency Issue Research Project, University of Waikato.
Linn, R.L. (1993) Linking results of distinct assessment. *Applied Measurement in Education* 6 (1), 83–102.
Maxwell, G. and Gipps, C. (1996) Teacher assessments of performance standards: A cross-national study of teacher judgements of student achievement in the context of national assesment schemes. Application for funding to the ARC: Interdisciplinary and International Research.
McGaw, B. (1996) Technical issues in assessments. Paper to the American Educational Research Annual Meeting, New York, April.
Mislevy, R'J. (1992) *Linking Educational Assessments. Concepts, Issues, Methods and Prospects*. Priceton, NJ: ETS.
QCA/DFEE *Assessment and Reporting Arrangements 1998*. London: QCA Publications.
Radnor, H. and Shaw, K. (1994) Developing a collaborative approach to moderation: The moderation and assessment project-southwest. In H. Torrance (ed.) *Evaluating Authentic Assessment*. Buckingham: Open University Press.
SCAA (1995a) *Consistency in Teacher Assessment, Key Stages 1 to 3*. London: SCAA.
SCAA (1995b) *Exemplification of Standards, English Key Stages 1 and 2, Levels 1 to 5*. London: SCAA.
SCAA (1995c) *Exemplification of Standards, Mathematics Key Stages 1 and 2, Levels 1 to 5*. London: SCAA.
SCAA (1995d) *Exemplification of Standards, Science, Key Stages 1 and 2, Levels 1 to 5*. London: SCAA.
SSABSA (Senior Secondary Assessment Board of South Australia) (1988) *Assessment and Moderation Policy*. Information Booklet No. 2. South Australia: SSABSA.