REVIEW

December 2004

A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes

Review conducted by the Assessment and Learning Research Synthesis Group

The EPPI-Centre is part of the Social Science Research Unit, Institute of Education, University of London

AUTHORS

This work is a review of the Assessment and Learning Research Synthesis Group (ALRSG).

The author of this report is Wynne Harlen, who conducted the review with the benefit of advice from the members of the ALRSG and with the active participation of the members indicated below.

Institutional base: Graduate School of Education
University of Bristol
35 Berkeley Square
Bristol BS8 1JA

REVIEW TEAM MEMBERSHIP

Mr Robin Bevan *	Deputy Head Teacher, King Edward VI Grammar School, Chelmsford
Professor Paul Black	King's College, University of London
Professor Richard Daugherty *	University of Wales, Aberystwyth
Mr Pete Dudley	Special Project Director, Classroom Learning, National College of School Leadership, and member of AAIA
Dr Kathryn Ecclestone	University of Exeter
Professor John Gardner *	Queen's University, Belfast
Professor Wynne Harlen *	University of Bristol
Dr Mary James	University of Cambridge
Ms Pru Rayner	Link Inspector for Primary Education, Nottinghamshire
Professor Judy Sebba *	University of Sussex
Dr Gordon Stobart	Institute of Education, University of London

[*Denotes members who were actively involved at certain parts of this review]

ADVISORY GROUP MEMBERSHIP

Dr Steven Bakker	ETS International, The Netherlands
Dr Dennis Bartels	Director, President, TERC, Cambridge, MA, USA
Professor Lorrie Shepard	University of Colorado
Professor Eva Baker	Co-director, CRESST, University of California, USA
Dr Terry Crooks	Director, EARM, University of Otago, Dunedin, New Zealand
Professor Dylan Wiliam	Educational Testing Service

ACKNOWLEDGEMENTS

This review was carried out with funding from the DfES through the EPPI-Centre (Evidence for Policy and Practice Information and Coordinating Centre) and followed the methodology developed by the EPPI-Centre. The help and advice of the EPPI-Centre staff who were linked to the review, Dr David Gough and Ms Zoe Garrett, is gratefully acknowledged. The author would also like to thank Elspeth Collins, Keren Durant and Marilyn Tew, for help in keywording and extracting data from studies. Without their help, the review could not have been completed in the reduced time allowed for the work. The collaboration with Dr. Ruth Deakin Crick, enabling not only assistants but also ideas to be shared across review groups, has been invaluable.

LIST OF ABBREVIATIONS

AAIA	Association for Achievement and Improvement through Assessment
ACCAC	The curriculum and assessment authority in Wales (no direct correspondence in English with the acronym)
AERA	American Educational Research Association
‘A’ level	External examination normally taken at the age of 18 and commonly required for university entrance
ALRSG	Assessment and Learning Research Synthesis Group
ARG	Assessment Reform Group
ATL	Association of Teachers and Lecturers
BEI	British Education Index
CBM	Curriculum-based measurement
CCEA	Council for the Curriculum Examination and Assessment (Northern Ireland)
CRESST	Centre for Research on Evaluation, Standards and Student Testing
CSE	Certificate of Secondary Education
DES	Department of Education and Science (previous name of current DfES)
DfEE	Department for Education and Employment (previous name of current DfES)
DfES	Department for Education and Skills
ERSDAT	Educational Research in Scotland Database
ERIC	Educational Resources Information Centre
EPPI-Centre	Evidence for Policy and Practice Information and Coordinating Centre
ETS	Educational Testing Service
GASP	Graded Assessments Science Project
GCE	General Certificate of Education
GCSE	General Certificate of Secondary Education
Kg	Kindergarten – usually for children aged 4–5 in the UK, 5–6 in the USA
KS	Key Stage. Used to identify stages of school education England and Wales. KS 1: ages 5–7; KS 2, ages 7–11; KS 3, aged 11; KS 4, ages 14–16
LD	Level description: specifically the description of a level in the National Curriculum
LEA	Local Education Authority (term used in England)
NAPS	National Assessment in Primary Schools: an evaluation project
NC	National Curriculum: in England. The curricula and their titles are different in Wales, Northern Ireland and Scotland.
NCA	National Curriculum Assessment
NCTM	National Council of Teachers of Mathematics

NFER	National Foundation for Educational Research in England and Wales
NUT	National Union of Teachers
PACE	Primary Assessment and Curriculum Experience project
QCA	Qualifications and Curriculum Authority, overseeing the curriculum and assessment in England
REEL	Research Evidence in Education Library
SAT	Standard assessment task / test
SCAA	Schools Curriculum and Assessment Authority (predecessor of QCA)
SEAC	School Examinations and Assessment Council
SEN	Special educational needs
SoA	Statement of attainment
TA	Teachers' assessment
TAS	Teacher Assessment Scheme (Hong Kong)

This report should be cited as: Harlen W (2004) A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. In *Research Evidence in Education Library.* London: EPPI-Centre, Social Science Research Unit, Institute of Education.

TABLE OF CONTENTS

SUMMARY

Background

This fourth review by the Assessment and Learning Research Synthesis Group (ALRSG) was designed to complement the previous review, which concerned the reliability and validity of teachers' assessment for summative purposes (Harlen, 2004). These reviews take place at a time of continued interest, at the highest levels of policy in the United Kingdom (UK), in giving a greater role to assessment by teachers in summative assessment. Recent evidence of this interest is found in the Primary National Strategy from the Department for Education and Skills (DfES), the commissioning by Qualifications and Curriculum Authority (QCA) in England of a review of 'Experiences of Summative Teacher Assessment in the UK', the acceptance by the Welsh Assembly of the report of the Daugherty Assessment Review Group in Wales, the 'Assessment is for Learning' project in Scotland, and the draft proposals of the Tomlinson Working Group on 14–19 education reform.

Assessment by teachers has the potential for providing summative information about students' achievement since teachers can build up a picture of students' attainments across the full range of activities and goals. Other benefits claimed include: less pressure on students and teachers compared with external tests and examinations; greater freedom of teachers to pursue learning goals in ways best suited to their students; the potential for information about students' ongoing achievements to be used formatively, to help learning, as well as for summative purposes; and the avoidance of the negative impact of tests on students' motivation for learning revealed by the first ALRSG review (Harlen and Deakin Crick, 2002). However, there are concerns about teachers' summative assessment relating to possible interference with the relationship between teacher and students, teachers' workload, and the need to ensure the quality and reliability of the outcomes. Experience in systems where teachers' assessment is successfully used for summative assessment shows that these problems can be overcome and confirms that there are benefits for students and teachers. This review was carried out to bring research evidence to bear on these different claims and experiences.

Definition of terms

Assessment in the context of education involves deciding, collecting and making judgements about evidence relevant to the goals of the learning. How these processes are carried out depends on the purposes of the assessment. The term 'summative assessment' refers to an assessment with a particular purpose – that of providing a record of a student's overall achievement in a specific area of learning at a certain time. It is the purpose that distinguishes it from assessment described as formative, diagnostic, or evaluative, rather than any particular method of gathering information about students' performance.

Although teachers inevitably have a role in any assessment, the term 'assessment by teachers' (teachers' assessment, often abbreviated to TA) is used for assessment where the professional judgement of teachers has a significant role in drawing inferences and making judgements, as well as in gathering evidence for assessment. Thus the definition of assessment by teachers for summative purposes used in this review, as in the previous review, is that it refers to any activity in which teachers gather evidence in a planned and systematic way about their students' learning to draw inferences, based on their professional judgement, to report achievement at a particular time.

This excludes assessment where teachers gather information that is marked by others, or teachers' involvement in setting or marking external examinations or tests.

The review questions

Considerations of the policy and practice background to this review led to the identification of the main review question as:

What is the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes?

To achieve its aims the review addressed the subsidiary question:

What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes?

The findings are used to address the further question:

What are the implications of the findings for policy and practice in summative assessment?

The outcomes of the review are as follows:

- the production of a map of studies reporting on the impact of using teachers' assessment for summative purposes on students, teachers and the curriculum
- the identification of the implications of the findings for different user groups, principally practitioners and policy-makers
- publication of the full report and of short summaries for different user groups in the Research Evidence and Education Library (REEL)
- identification of further research that is needed in this area

Methods

The review methodology followed the procedures devised by the EPPI-Centre, with the technical support of the EPPI-Centre. Criteria were defined for guiding a wide-ranging search for studies that dealt with some form of summative assessment conducted by teachers, involving students in school in the age range 4 to 18 and reporting on the impact of the teachers' assessment on students, teachers or the curriculum. Bibliographic databases and registers of educational research were searched online as were relevant journals online, with other journals and back numbers of some of those available online being searched by hand. Other studies were found by scanning the references lists of already identified reports, making requests to members of relevant associations, other review groups, and using personal contacts.

All studies identified in these ways were screened, using inclusion and exclusion criteria, and the included studies were then keyworded, using the EPPI-Centre Core Keywording Strategy (EPPI-Centre, 2002a) and additional keywords specific to the context of the review. Keywords were used to produce a map of selected studies. Detailed data extraction was carried out online. using EPPI-Centre generic and review-specific data extraction guidelines (EPPI-Centre, 2002b). Two reviewers worked independently before comparing entries and reaching a consensus. Judgements were made as to the weight of evidence relevant to the review provided by each study in relation to methodological soundness, appropriateness of the study type and relevance of the focus to the review questions.

The structure of the synthesis was based on the impact reported by the studies. Potentially there were three main headings, bringing together the findings for impact on students, teachers and the curriculum. However, since few studies were found that dealt with the curriculum – and those that did were also concerned with impact on teachers – these were combined into one group of impact on teachers and the curriculum. Within these two broad groups, formed according to the nature of the impact reported, there are important subdivisions according to whether the assessment is used for internal school purposes only (such as grades and routine school tests and examinations) or for use by others outside the school (as in the case of certification, selection, transfer, or the accountability of the school). Although it was often the case that an assessment was used for both internal and external purposes, it was considered helpful to try to identify the main use and to discuss studies under these subheadings.

Potential users of the review, represented in the Review Group, were involved in several ways: providing advice during and between Group meetings; providing information about studies; participating in keywording and data extraction; and commenting on draft findings and implications to be drawn from them.

The ALRSG includes the following users: a secondary school deputy head teacher with responsibility for assessment, a local authority primary adviser and a project director of the National College of School Leadership. Two members of the group are members of AAIA, another led the review of assessment in Wales and another is Director of the Learning to Learn project of the ESRC's Teaching and Learning Research programme. Eight of the Review Group are members of the Assessment Reform Group and, through this, the Review Group has an ongoing relationship with the DfES and with the QCA.

Results

Identification of studies

The search for studies resulted in 343 papers being found, of which 301 were excluded in either a one-stage or a two-stage screening process, using inclusion and exclusion criteria. Full texts were obtained for 26 of the remaining 42 papers, from which a further two were excluded during keywording. One paper was found to be linked to another and one of these papers was then excluded as a separate item, as it was based on the same set of data. This left 23 studies after keywording. All these were included in the systematic map and in-depth review.

Systematic map

The 23 studies included in the in-depth review were mapped in terms of the EPPI-Centre and review-specific keywords. All were written in the English language; 12 were conducted in England, nine in the United States and one each in New Zealand and Hong Kong.

All studies were concerned with students between the ages of 4 and 18. Eleven involved primary school students (aged 10 or below) only, six involved secondary students (aged 11 or above) only, and five were concerned with both primary and secondary students. A slightly larger proportion of studies conducted in primary schools reported impact on teachers compared with those conducted in secondary schools. About 70% of studies in secondary schools and about 80% in primary schools were concerned with assessment of English; while 43% and 60% respectively were concerned with assessment of mathematics.

Twenty studies were classified as involving assessment of work as part of, or embedded in, regular activities. Three were classified as portfolios, two were classified as projects and eight were either set externally or set by the teacher to external criteria. The most common use of the assessment in the studies was for internal school purposes, with four studies related to assessment for certification and another three to external purposes that had high stakes for the school.

In-depth review and synthesis

Seven of the 23 included studies provided evidence of high weight for the review. Six of these provided information about impact on students; three also provided information about impact on teachers. Of the 12 studies providing evidence of medium weight, all except one provided evidence of impact on teachers, whilst five provided information of impact on students.

Findings from studies relating to impact of teachers' assessment on students

When teachers' assessment is used for *external* purposes, there was high weight evidence of the following:

- Older students respond positively to summative assessment by teachers of their coursework, finding the work motivating and being able to learn during the assessment process (Bullock *et al.*, 2002).
- Students need more help, in the form of better descriptions and examples, to understand the assessment criteria and what is expected of them in meeting these criteria (Bullock *et al.*, 2002; Iredale, 1990; Stables, 1992).
- The impact of summative teacher assessment on students depends on the high stakes use of the results (Yung, 2002).
- The impact of summative teachers' assessment on students will be affected by the way teachers interpret their roles as assessors and by their orientation towards improving the quality of students' learning or maximising their marks (Bullock *et al.*, 2002; Yung, 2002).

There is medium weight evidence of the following:

- Teachers consider that young students may not do their best work when constrained by an external task (Abbott *et al*., 1994).

When teachers' assessment is used for *internal purposes,* there is high weight evidence in relation to impact on students as follows:

- Feedback from earlier assessment impacts on the effort that students apply in further tasks of the same kind; effort is motivated by non-judgemental feedback that gives information about how to improve (Brookhart and DeVoge, 1999; Carter, 1997/8).
- The way in which teachers present classroom assessment activities may affect students' orientation to learning goals or performance goals (Brookhart and DeVoge, 1999).
- Changing teachers' assessment practices to include processes and explanations can lead to better student learning (Flexer *et al*., 1995).
- Using grades as rewards and punishments is harmful to students' learning by encouraging extrinsic motivation (Iredale, 1990; Pilcher, 1994).

There is medium weight evidence of the following:

- Teachers' own unguided grades are influenced by non-achievement factors, such as students' behaviour, effort, attendance and disadvantaging some students (Bennett *et al.*, 1993; Cizek *et al.*, 1995/6).
- The introduction of teachers' assessment related to levels of the National Curriculum in England and Wales was perceived by teachers as having a positive impact on students' learning experiences (Hall *et al*., 1997).

Findings from studies relating to impact of teachers' assessment on teachers and the curriculum

When teachers' assessment is used for *external* purposes there is high weight evidence of the following:

- Teachers vary in how they respond to being given the role of assessor and the approach they take to interpreting external assessment criteria; strict adherence to the regulations leads them to be less concerned with students as individuals (Morgan, 1996; Yung, 2002).

There is medium weight evidence of the following:

- The impact on teaching of external assessment requirements depends on the value that teachers find in the information they gain about their students through the assessment (Abbott *et al*., 1994; Bennett *et al*., 1992; Koretz *et al*., 1994).
- Assessment for external purposes adversely affects teachers when it is seen as taking up too much time from teaching (Abbott *et al*., 1994; Bennett *et al*., 1992).

When teachers' assessment is used for *internal* purposes, there is high weight evidence in relation to impact on teachers and the curriculum as follows:

- The introduction of assessment techniques that require students to think more deeply leads to changes in teaching that extend the range of students' learning experiences (Flexer *et al*., 1995).
- Close external control of teacher assessment inhibits teachers gaining detailed knowledge of their students (Johnston *et al*., 1993).

There is medium weight evidence of the following:

- When teachers' assessment is built into teachers' planning, the process has a positive impact on teaching and learning. This impact is further enhanced by professional collaboration at the school level (Hall *et al*., 1997; Hall and Harding, 2002).
- Assessment by teachers indicates where learning opportunities for their students need to be extended (Valencia and Au, 1997; Whetton *et al*., 1991).
- In a low stakes context, the process of summative assessment by teachers helps them to clarify the meaning of learning outcomes (Valencia and Au, 1997).
- The value of teachers' summative assessment of potential users depends on teachers internalising the nature of progression in relation to the learning goals (Cizek *et al*., 1995/6; Hill, 2002).

Findings in relation to the conditions and contexts affecting the nature and extent of the impact of using teachers' assessment for summative purposes

There is both high and medium weight evidence of the following:

- New assessment practices are likely to have a positive impact on teaching if teachers find them of value in helping them to learn more about their students and to develop their understanding of curriculum goals; time to experience and develop some ownership of practices enhances their positive impact (Abbott *et* *al*., 1994; Bennett *et al.,* 1992; Flexer *et al*., 1995; Gipps and Clarke, 1998; Koretz *et al*., 1994).

- When high stakes judgements are associated with teachers' assessment, one effect is for teachers to reduce assessment tasks to routine events and restrict students' opportunities for learning from them; high stakes encourages some teachers to give high grades where there is doubt, which may not be in the students' interests (Bullock *et al*., 2002; Hall and Harding, 2002; Morgan, 1996; Yung, 2002).

- Shared criteria for assessing specific aspects of achievement lead to positive impact on students and on teaching; in the absence of such guidance, there is little positive impact on teaching and a potential negative impact on students (Bennett *et al*., 1993; Cizek *et al*., 1995/6; Hall *et al*., 1997; McCallum *et al*., 1993; Pilcher, 1994).

- The process that teachers use in setting assessment tasks and in grading impacts on students' motivation for learning, particularly their goal orientation, when grades are used as rewards or punishments; the negative impact can be alleviated by ensuring that students have a firm understanding of assessment processes and criteria (Brookhart and DeVoge, 1999; Bullock *et al*., 2002; Iredale, 1990; Stables, 1992).

- Summative assessment by teachers has a more positive impact on teachers and teaching when integrated into practice than when concentrated at a certain occasion (Bennett *et al*., 1993; Bullock *et al*., 2002; Carter, 1997/8; Hall *et al*., 1997; Iredale, 1990; Johnston *et al*., 1993; Koretz *et al*., 1994; McCallum *et al*., 1993; Whetton *et al*., 1991).

- Opportunities that enable teachers to share and develop their understanding of assessment procedures enables them to review their teaching practice and their view of students' learning and of subject goals; such opportunities need to be sustained over time and should preferably include provision for teachers to work collaboratively across as well as within schools (Flexer *et al*., 1995; Gipps and Clarke, 1998; Hall *et al*., 1997; Hall and Harding, 2002; Hiebert and Davinroy, 1993; Valencia and Au, 1997).

Conclusions

Strengths and limitations of the review

The strengths of the review emanate from its systematic and collaborative procedures. The documentation of searches and of inclusion and exclusion decisions enables the work to be extended at a later date without duplication. All critical decisions about inclusion, exclusion and weight of evidence were taken by at least two people working first independently and then reconciling any differences in judgements. The main limitations in relation to procedures arise from the search being confined to studies published in English and available either online, in the university library or via inter-library loan. The findings are limited to some extent by the small number of studies found that provided evidence of high weight in relation to the review questions.

Implications for policy

- Summative assessment by teachers has the potential for positive effects on students and on teachers, without the negative effects associated with external tests and examinations.

- Using teachers' assessment for summative purposes can support valid assessment of key learning processes as well as assessment of learning outcomes related to higher level cognitive skills.

- Summative assessment by teachers has most benefit when teachers use evidence gathered over a period of time and with appropriate flexibility in choice of tasks rather than from an event taking place at a particular time. This enables information to be used formatively to adapt teaching as well as summatively.

- Using the results of student assessment for high stakes school accountability reduces the validity of the assessment, whether this is conducted by teachers or by external tests and examinations.

- Introducing new assessment practices can support beneficial change in teaching, providing that the techniques are well matched to learning goals and illustrate how students can be required to use important conceptual knowledge and leaning skills.

- Regulations for teachers' summative assessment should allow teachers opportunities to assimilate summative assessment into their practice and to design appropriate classroom programmes. When changes are made in assessment practices, time must be allowed for this assimilation to happen.

Implications for practice

The following actions are likely to increase the benefit of teachers undertaking summative assessment of their own students:

- At all stages and for all purposes, students should be helped to understand the criteria by which their work is assessed. This is likely to mean providing and discussing examples that illustrate the practical meaning of the criteria.

- Teachers should make explicit to all concerned – colleagues, parents and students – the basis of the marks and grades they assign for internal school purposes. Achievement grades should not be influenced by non-academic factors, such as behaviour and participation, which should be reported separately as appropriate.

- When presenting assessment tasks to students, teachers should emphasise learning outcomes and not the attainment of a high grade, thus avoiding the encouragement of extrinsic motivation which leads to shallow learning.

- Teachers should internalise the progression in skills and understanding they aim to help students develop and interpret student performance in these terms rather than use a checklist of specific unconnected behaviours. In this way summative assessment helps teachers' understanding of learning goals as well as facilitates more detailed knowledge of their students.

- Schools should set aside time for teachers to discuss assessment issues, plan assessments and moderate their judgements of students' work. This not only improves the reliability of the assessment but enables teachers to use the process of summative assessment to help teaching and learning.

Implications for research

The low number of studies found that met the inclusion criteria for this study, with only seven providing evidence of high weight, leads to an obvious implication that more research and more high quality research is needed in this area. Given the interest at high levels in government in making greater use of teachers' assessment in summative assessment, indicated in the Background section at the beginning of this chapter, there is some urgency in meeting the need for more research.

Particular research foci suggested by this review are as follows:

- How teachers manage the dual roles as teacher and assessor
- The impact on students and on other uses of assessment of changing from tasks devised and marked externally to using teachers' judgements of students' performance in special tasks and in regular work
- The identification of factors that support teachers use of summative assessment to improve students' learning experience: that is, how the formative use of assessment can be integrated with the summative use
- Direct comparison of different approaches used by teachers in summative assessment to investigate whether they make any difference to outcomes or to impact on students
- Investigation of what information is actually used by teachers in their assessment and what impact this has on the curriculum experience by students
- The role of student self-assessment in summative assessment
- The impact on students of developing their awareness of success criteria and providing exemplification of learning goals
- What changes to accountability procedures would preserve the integrity of teachers' assessment and minimise pressures to give inflated grades or levels

1. BACKGROUND

This chapter begins by summarising the events leading to the proposal for this systematic review, the fourth conducted by the Assessment and Learning Research Synthesis Group (ALRSG). The review is the second one that the group has undertaken on the subject of summative assessment by teachers and complements the third review, which concerned the evidence pertaining to the reliability and validity of teachers' assessment for summative purposes. This review takes place at a time of continued interest at the highest levels of government and its agencies concerned with the school curriculum and assessment in giving a greater role to assessment by teachers in summative assessment. Recent evidence of this interest is found in the Primary National Strategy from the DfES, the commissioning by QCA in England of a review of 'Experiences of Summative Teacher Assessment in the UK', the acceptance by the Welsh Assembly of the report of the Daugherty Assessment Review Group in Wales, the *Assessment is for Learning* project in Scotland, and the draft proposals of the Tomlinson Working Group on 14-19 reform.

The chapter continues with a discussion of the meaning being given here to terms, such as 'summative assessment' and 'teachers' assessment'. Later sections concern the policy, practice and research background to the review, and the identification of the review questions.

1.1 Aims and rationale

1.1.1 Previous work of the ALRSG

The ALRSG was created as one of the first wave of EPPI-Centre Review Groups in 2000 and undertook its first review from February 2001 to January 2002. This was entitled 'A systematic review of the impact of summative assessment and testing on students' motivation for learning' and was published in the Research Evidence in Education Library (REEL) in 2002 (Harlen and Deakin Crick, 2002). The second review, conducted from February 2002 to January 2003, was concerned with the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills, published on REEL in 2003 (Harlen and Deakin Crick, 2003a).

In February 2003, the group embarked on a two-year plan of review work focused on the use of assessment by teachers for summative purposes. This focus was in response to evidence from previous reviews (Crooks, 1988; Black and Wiliam, 1998) that, on the one hand, formative assessment can raise standards of achievement, and, on the other, that claims that testing raises achievement are false (Flexer *et al*., 1995; Koretz *et al*., 1991; Linn, 2000). At best, repeated testing raises test scores but without any real improvement in achievement. Further, the first ALRSG review found that testing has a negative impact on motivation for learning. To avoid these negative effects of external testing, there were indications that policy-makers' attention was turning to considering greater use of assessment by teachers as an alternative to testing. There were, however, concerns about the dependability and effects of assessment by teachers used for summative purposes. The third and fourth reviews were therefore set up to identify the evidence base in relation to these concerns. The third review, on the reliability and validity of assessment by teachers for summative purposes, was completed in February 2004 and published in March 2004. The fourth review is concerned with the impact that using teachers' assessment, as all or part of summative assessment, has on students, teachers and the curriculum.

1.1.2 Rationale

Claims are made that assessment by teachers for summative purposes holds the promise of:

a. reducing the pressure on teachers and students from external tests and examinations

b. enabling teachers greater freedom to pursue and assess their own goals

c. providing formative feedback to students through being conducted as part of teaching, as well as providing information for summative assessment (Crooks, 1988)

However, in practice, there are problems of ensuring these benefits. First, teachers feel pressure of a different kind in being both teacher and assessor, a dual role which some suggest interferes in relationships with students (Morgan, 1996). Second, there is concern about the additional time required for making and recording assessment and for the moderation processes that are required when the outcome is for 'external' use. Third, unless there is effective professional development in the processes of assessment, teachers fall back on familiar methods and emulate the form and scope of external tests in their own assessments; there is also evidence that these teacher-made tests are of low quality (McMorris and Boothroyd, 1993). Fourth, there is a considerable degree of mistrust of assessments based on teachers' judgements.

Counters to these concerns are that they occur when changes are made without proper preparation of teachers and of users of assessment. In successful implementation (as in Queensland, Australia, as noted below), teachers are involved in deciding the programme of work and have some ownership of the assessment scheme. Even without this degree of involvement, there is some evidence that having a central role in assessment sharpens teachers' understanding of the learning objectives and focuses their teaching on all the objectives, rather than on those that are tested by external examinations (Frederiksen and Collins, 1989; Koretz *et al*., 1994; NCEST, 1992). Providing that the process is an open one, in which students are aware of what they are aiming for and how it will be assessed, there is no need for damage to the teacher-student relationship.

There is clearly a need to bring evidence to bear in relation to these different claims and experiences. Considerable importance attaches to the consequences of assessments (Messick, 1989) and, as Linn (1994) has pointed out, it is not sufficient to show that an assessment has construct validity. 'Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and, at the same time not producing unintended negative outcomes'

(Linn, 1994, p 8). It was the purpose of this review to identify implications for policy and practice, by studying the circumstances in which the advantages of using TA for summative purposes can be achieved without too many of the disadvantages.

1.2 Definitional and conceptual issues

1.2.1 Educational assessment

Assessment in the context of education involves deciding, collecting and making judgements about evidence relevant to the goals of the learning being assessed. There is a wide range of ways of gathering evidence; which of these is chosen in a particular context depends on the purposes of the assessment. Making judgements involves considering the evidence of achievement of the goals in relation to some standards, or criteria or expectations. Again, how this is done will depend on the purpose, so this is a key factor to take into account.

Pophams' definition of educational assessment could apply to formative or summative purpose but assumes an active role of the teacher in the process 'by which teachers use learners' responses to specially created or naturally occurring stimuli to draw inferences about the learners' knowledge and skills' (Popham, 2000, quoted in NRC, 2001, p 20).

1.2.2 Summative assessment

The term 'summative assessment' refers to an assessment with a particular purpose – that of *providing a record of a student's overall achievement in a specific area of learning at a certain time*. It is the purpose that distinguishes it from assessment described as formative, diagnostic, or evaluative (DES, 1987). Thus a particular method for obtaining information, such as observation by teachers, could, in theory, be used for any of these purposes and so does not identify the assessment as formative, summative, etc. Consequently, in this discussion of the use of teachers' assessment for summative purposes, it is important to keep in mind the distinction between purposes and methods of gathering information for assessment.

1.2.3 Teachers' assessment

Although teachers inevitably have a role in any assessment, the term 'assessment by teachers' (teachers' assessment, often abbreviated to TA) is used for assessment where the professional judgement of teachers has a significant role in drawing inferences and making judgements as well as in gathering evidence for assessment. Teachers may use observation during regular activities, or set up special tasks or projects to check what students can do or what ideas they have; alternatively, they may use classwork, coursework or short tests that they construct themselves. In setting these tasks and drawing inferences from the outcomes, they are comparing outcomes with some standard or expectation. Even in the most informal approaches, teachers will be seeking evidence in relation to particular learning goals, which will frame and focus their attention. In more formal approaches, they may be using criteria or even checklists developed by others.

In some school-based assessments teachers have a role only in gathering evidence that is then marked or graded by others. Since the process does not involve the students' own teachers in using their professional judgement, assessment of this kind is not included in the meaning of *teachers' assessment* or *assessment by teachers* used in this review.

There is a widespread assumption that teachers' assessment serves a formative function, whilst externally produced tests or other assessment procedures serve a summative function. However, this is not by any means always the case. Whilst a truly formative assessment can only be based on teachers' assessment, the fact that a teacher makes decisions about and conducts an assessment does not necessarily mean that it serves a formative function. The key test of whether the assessment is or is not formative is whether or not the findings are linked to teaching and learning: that is, the extent to which it provides some information that the teacher needs and uses to help the students learn. In summative assessment, this use is not a requirement since the purpose is primarily to report on learning to the various stakeholders – students, parents, other teachers, employers, assessment agencies, etc.

The definition of assessment by teachers for summative purposes used in this review, as in the previous review, is that it refers to any activity in which teachers gather evidence in a planned and systematic way about their students' learning to draw inferences based on their professional judgement to report achievement at a particular time.

1.2.4 Other terms used to describe assessment by teachers

Performance assessment is a term, used mainly in the United States of America (USA), to mean an assessment that requires mental processes and physical actions that more closely reflect the goals of learning than standardised tests. These assessments can involve students in problem solving, practical tasks, projects, extended writing, presentations, etc., and in the collection of information through methods such as interviews and observation. The rationale is that not only do they provide more valid student assessment but 'that they serve as motivators in improving students' achievement and learning, and that they encourage instructional strategies and techniques that foster reasoning, problem solving and communication' (Lane *et al.*, 2002, p 280). In other words, if teachers teach to these tests, this will not be as damaging as teaching to standardised tests (Shepard, 1990, p 21).

A term used interchangeably with performance assessment is *authentic assessment,* which Torrance (1995) describes as implying that 'assessment tasks designed for students should be more practical, realistic and challenging than what one might call 'traditional' paper-and-pencil tests' (Torrance, 1995, p 1). *Embedded assessment* is a term that implies not only that the assessment tasks are very similar to regular learning activities, but that they are combined with them so that students are scarcely aware of being assessed.

In the UK, the terms *school-based assessment* and *coursework assessment* are used to describe assessment that take place in the school. It is not necessarily embedded and some school-based assessment includes tests created by the students' own teachers. Further, coursework is not always assessed by the students' own teachers and so would not invariably fall into the definition of TA used here.

All these types of TA combine some degree of formative purpose with a summative purpose. This is even more so in the case of *dynamic assessment.* This is a term used to describe assessment by teachers which aims to stretch students by giving them tasks a little beyond their present level. Help is provided if necessary, but the purpose is to see what students can achieve on their own (Brown *et al*., 1993). *Continuous assessment* and *ongoing assessment* are ambiguous terms, sometimes used to refer to TA for formative purposes, but sometimes to mean that a continuous record of achievement is kept as a basis for summative assessment at required times.

1.3 Policy and practice background

1.3.1 Policy

All teachers who report to their students' parents, or provide records for other teachers, are involved in assessment for summative purposes. The reports and records required vary in structure, detail and form according to the school, local authority, the national context and the age of the students. However, these records serve what might be called 'internal' purposes of assessment: that is, they are for the information of those primarily concerned with helping the further learning of the students. In contrast, summative assessment that has 'external' uses – that is, results are reported to the authorities outside the school – may be used for certification or selection that can make a difference to students' further opportunities, or be used for accountability purposes in relation to teachers or schools. The important consequences of external assessment mean that it is higher stakes than internal assessment.

An example of internal summative assessment by teachers is the statutory requirement in England for assessment in certain subjects at the end of the Foundation Stage and at the end of each Key Stage. In all cases, there is a component of assessment by teachers; at Foundation Level all assessment is by teachers. In addition, all schools are required to report students' achievements to parents at least once every year. For non-end-of-Key-Stage Years, this is on the basis of the teachers' assessment (TA), although there are some optional, internally marked, tests which teachers can use to help their judgements of the levels achieved. Since TA is not used in creating league tables of schools, it does not have the high stakes that attach to the national tests, which are used for this purpose.

The extent to which the summative teachers' assessment (TA) is likely to have an impact on students, teachers or the curriculum depends on the degree of formality, the required mode of reporting or recording judgements and the level of 'stakes' attached to the result. For instance, in England, Foundation Level assessment is low stakes, although statutory. The requirements are to record progress in relation to 13 scales within six areas of learning, with a final profile being completed in the last term of the Foundation Stage. All assessments are made within the normal course of activities and are intended to serve formative, as well as summative, purposes. Thus any impact on teachers is likely to be through the focusing effect of the scales and the associated guidance for practitioners.

An alternative situation arises, generally with older students, when teachers' assessment of special tasks or projects is used in whole or in part for certification purposes. This was the case with the early GCSE examination and its predecessors, the CSE and GCE. The intention is to ensure that parts of the subject that cannot be assessed through external written tests, such as practical science and spoken foreign language, are included. Distrust of teachers' judgements led, in 1992, to the government in England and Wales limiting the proportion of credit that could be awarded on the basis of assessment by teachers.

However, assessment by teachers continues to be a component of summative assessment for certification in many countries, including Sweden, the Australian States of Queensland and Victoria, the Caribbean and the UK (Black, 1998; Broadfoot *et al.*, 1990; Maxwell, 1995, 2004; Wood, 1991). It is widely used as the only form of assessment for many post-graduate courses and for vocational and professional certification. The modularisation of courses, where each unit or module is separately assessed by teachers, theoretically enables assessment to have a formative as well as a summative role. However, this continuous assessment accentuates the possible conflict in roles when the teacher is both the supporter or provider of learning and the judge of the achievement. This is particularly so when the assessment has high stakes (Choi, 1999).

There have recently been indications of willingness among policy-makers to consider a greater role for teachers in summative assessment for external as well as internal purposes. This is partly driven by concerns that have been raised about the effect of tests on students' motivation for learning (e.g. ARG, 2002; Harlen and Deakin Crick, 2003b), on students (Abbott *et al.*, 1994; Pollard *et al.*, 2000) and on the curriculum (Crooks, 1988; James, 1998; Shorrocks *et al.*, 1992; Wood, 1991). Further, in the light of evidence (Black and Wiliam, 1998) of the role that formative use of assessment can have in raising levels of achievement, the evidence that the practice of formative assessment was declining in the face of pressures due to external tests (Pollard *et al.*, 2000) has added to the impetus to consider alternatives to testing. Thus, the Chief Inspector of schools in England has recognised that, rather than use test scores as indicators of schools' achievements, 'the time is now right to take greater account of what head teachers are saying about the students in their own school, and, more specifically, what strategies they will deploy to improve attainment' (Bell, 2003). The QCA has undertaken a review of teacher assessment models which have been implemented in the UK since 1950. Further, the QCA report on comparability of national tests over time (Massey *et al.*, 2003), noted the following:

> …teacher assessment showed less sign of drifting standards than national tests in Reading/English and Mathematics. Teacher assessment appears in this light less unreliable than might have been assumed when the current national testing system was designed. (p 239)

Thus there is interest at the policy level in ways of using TA for summative assessment but it is clearly important to be aware of any possible impacts that can be identified from existing practice in the UK and other countries.

1.3.2 Practice

Despite the fact that conducting and reporting their summative assessment of students has always been an established part of teachers' roles, no particular attention was paid to its impact until the introduction of the National Curriculum Assessment (NCA) in England and Wales, and similar innovations in Northern Ireland and Scotland in the early 1990s. The NCA introduced into primary schools, where teachers' summative assessment had been much less obvious than in secondary schools, a new aspect of the teachers' role, as an assessor, that was perceived to be in conflict with the role of facilitating learning. Gipps (1994), for example, states that

> Where school-based teacher assessment is to be used for summative purposes then the relationship between teacher and student can become strained: the teacher may be seen as a judge rather than facilitator. This uneasy dual role for the teacher which ensues is a result of the formative/summative tension. If the teachers' assessment were not to be used for summative purposes then the relationship could stay in the supportive mode (p 127).

Two extended projects followed the course of the introduction of the NCA in primary schools in England. The National Assessment in Primary Schools: an Evaluation (NAPS) project followed teachers from 1990 to 1994. During this time, teachers were required to assess their students against National Curriculum attainment targets, or, later, against level descriptions and arrive at a 'level'. Before 1993, TA was to be combined with the national test result to provide an overall level; after that date, the two were to be reported separately. Since little guidance was given to teachers on how to conduct TA – most attention being given to the development, trial and implementation of national tests – teachers devised their own procedures, which were researched in the NAPS project. The researchers found a wide range of different ways of collecting information and of relating it to National Curriculum levels, which they grouped into three models of teachers' assessment (McCallum *et al.*, 1993; Gipps *et al.,* 1995). Although these models grew out of different teaching styles, they noted that the requirement for overt TA had had an impact on teachers' ways of working. Some head teachers judged this impact to be positive, making teaching more focused and teachers more aware of what students should be achieving. Others were concerned about the pressures on teachers and the effect of the national tests in narrowing the curriculum. Teachers themselves reported changes in what they taught as a result of the introduction of the national tests. They also recognised changes in their teaching behaviour – for instance, in questioning, doing more observation and in making notes of events that were evidence of students' achievement.

The Primary Assessment, Curriculum and Experience (PACE) project looked at the impact of the introduction of NCA on students. They reported some negative impact, although it was not possible to disentangle an impact due to TA from that due to the national tests. For whatever reason, the project found the following:

> As Key Stage Two progressed, the children's feelings of anxiety developed further as teachers increased the amount of routine testing. Additionally they often felt uncertain and vulnerable when ambiguous classroom tasks were combined with a high-stakes, categoric assessment climate (Pollard *et al.*, 2000, p 285).

They also concluded that 'for these students, assessment had more to do with pronouncing on their attainment than with progressing their learning' (ibid), clearly implying that more TA for summative purposes had reduced TA for formative purposes.

The NCA is high stakes for the teachers rather than for the students. In cases where the TA is all or part of assessment for an external award – high stakes for the student – there is not only the problem of the dual role of the teacher, but there can also develop 'in students the mindset that if a piece of work does not contribute towards the total, it is not worth doing' (Sadler, 1989, p 141). The development of this mindset is particularly common in the context of modular courses, where students feel constantly under scrutiny.

However, no such problems have been reported in relation to the school-based assessment scheme for awarding the Senior Certificate that has been in place in Queensland, Australia in some form since 1971. At first it was a norm-based assessment, but it was converted to a criterion-based scheme in 1981. What makes this different from TA that is dictated from outside the school, is that the schools in Queensland are responsible for their own 'work programme', which sets out objectives, course content and the assessment plan. Thus they have ownership of decisions about assessment. The work programme is regularly updated and accredited by the Examination Board and is publicly available.

> The school work programme is an important document in the criterion and standards referenced system of assessment in Queensland. The work programme is usually placed in the school library and can be consulted by the students or the parents. The specific objectives are often given to the students for each syllabus topic so that they are clear about what has to be learnt and the standards of achievement required. Knowledge of objectives gives the students the power to manage their own learning and to check on the completeness of the treatment of the syllabus topic by their teacher (Butler, 1995, p 144).

Thus there is involvement of teachers in all parts of the procedures of creating the school programme for the Senior Certificate, implementing procedures and applying criteria to documented student performance. Further, the openness of the procedures, particularly the sharing with students, avoids creating anxiety through uncertainty about what is required of them. Butler (1995) notes that the scheme has continued with no major problems or disruptions 'from the Board, the politicians, the teachers or the public' (p 153). He also notes that this system 'is much less costly than state-wide external examinations' (ibid). A system of local Review Panels maintains comparability of standards across the state.

Maxwell (2004), poses the question: 'What is needed to make such an approach successful?' and answers it as follows:

> Foremost, it is necessary to believe that teachers can acquire the appropriate expertise and that they will act professionally and ethically. Certainly, a premium is placed on assessment expertise. However, the need for teachers to become skilled in conducting assessment programs and judging the quality of students' performance against defined assessment standards creates its own impetus for teachers to acquire these skills. Teachers typically take up the challenge when they are given the responsibility. (Maxwell, 2004, p 6)

Thus there are lessons to be learned from schemes that appear to be successfully using TA for summative purposes, even where high stakes are attached to the outcomes. A significant aspect is the involvement of teachers in decisions about what to assess, which brings not only commitment but understanding of what and how to assess. This is likely to counter the tendency reported by several researchers (Black, 1993; Choi, 1999; Lubisi and Murphy, 2002) for teachers to emulate external tests in their own assessments.

1.4 Research background

This review is closely related to the most recent review conducted by the ALRSG, on the evidence of the reliability and validity of assessment by teachers used for summative purposes (Harlen, 2004). Some of the studies included in that review provided information about impact in addition to reporting evidence relating to reliability and/or validity. For example, Koretz *et al.* (1994) found that a portfolio system for reporting students' achievement had the desired effect in relation to the programme's goal of improving the range and nature of activities provided by teachers. However, portfolio systems have been found to have low reliability and validity. Hall *et al.* (1997) reported that the introduction of TA in the National Curriculum assessment of 7-year-olds in England and Wales caused teachers to plan in greater depth, although it had a less positive impact by increasing concentration on curriculum coverage at the expense of following their own or students' interests. Radnor *et al.* (1995), Shorrocks *et al.* (1992), Gipps *et al.* (1996) and Abbott *et al.* (1994) all provide some evidence of impact, not exploited in the previous review. However many studies excluded from the previous review provide information relevant to impact.

Brookhart (1994) reviewed research on teachers' grading practices. She concluded that classroom assessments have profound effects on students (p 291) and placed emphasis on their motivational impact, particularly in relation to conation (will). She also referred to the use of grades as a management tool on account of their importance both within the classroom and outside where they may be linked to various rewards, including parental approval. Carter (1997/8) reported a positive impact on high ability students of being given responsibility for analysing their own test papers.

There are many claims that involving teachers as markers or graders of classroom tests, even if they are not assessing their own students, has a positive impact on their understanding of performance-based learning and teaching. Gilmore (2002) reported a positive impact of this experience, using evidence from teachers' perceptions of change in their confidence and understanding in relation to assessment. However, in a study of teachers involved in grading the Maryland School Performance Assessment Program (MASAP), Goldberg and Roswell (1999) examined actual classroom practice. They found little evidence of real change in practice, despite a greater understanding of key aspects of the programme by teachers who had been involved in grading compared with those who had not. Whilst noting that 'teachers almost universally perceive scoring as a valuable learning experience' (p 287), they suggest that researchers should 'move beyond the anecdotal, not only to the examination of classroom artefacts but to the context in which those artefacts are created and used' (ibid).

Different ways in which teachers interpret regulations for classroom-based assessment that contributes to examination grades was the subject of a study by Yung (2002). The extent to which the Teacher Assessment Scheme (TAS) introduced in Hong Kong has 'a liberating influence on the curriculum and would bring about a host of desirable curricular and pedagogical changes' (p97) was found to vary according to teachers' beliefs, professional confidence and consciousness. Some lack of confidence was considered to be the result of teachers previously being treated as technicians and subject to bureaucratic accountability. This echoes the earlier report of Donnelly *et al.* (1993) that external moderation can lead to a loss of professional autonomy, with teachers concerned about 'passing' the moderation.

1.5 Authors, funders and other users of the review

This review is the fourth EPPI-Review carried out by the Assessment and Learning Research Synthesis Group (ALRSG). Current members of the Review Group and overseas advisers are listed in Appendix 1.1. The review was proposed and conducted because of evidence, revealed by the first ALRSG, among other sources, of the negative impact of tests on students' motivation for learning and on account of the recent interest in alternatives to testing for summative assessment, in the form of assessment by teachers.

The author of this report is Wynne Harlen, based at the Graduate School of Education of the University of Bristol, where she is a Visiting Professor in Education. The review was funded solely by the contract between the EPPI-Centre at the Institute of Education, and the University of Bristol on behalf of the ALRSG. The review was carried out by the author with the guidance of the ALRSG and participation of its members, including teacher and adviser members, at various stages as noted in section 2.1. The ALRSG includes all members of the Assessment Reform Group (ARG), a voluntary group of researchers who have, since 1989, worked to ensure that research in assessment is used to inform policy and practice in educational assessment. During 2003, the ARG was awarded a grant by the Nuffield Foundation to conduct a series of expert seminars, spread throughout 2004/05, on the topic of 'Assessment systems of the future: the place of assessment by teachers'. The findings of the third review were the main input into the first seminar in the series, attended by policy-makers, advisers and teachers from all parts of the UK. The project will also provide a platform for consultation with users on the outcomes of the current review.

1.6 Review questions

1.6.1 Aims

The arguments that summative assessment needs to reflect the full range of learning goals, and that these goals include a number of learning outcomes that are not well suited to assessment by conventional tests, support the case for giving assessment by teachers a greater role in summative assessment in addition to their role in formative assessment. There is also strong evidence that conventional tests have negative impacts on students' motivation for learning, on teachers' methods and on the curriculum. A further reason is that testing has a limiting effect on the practice of formative assessment, which is known to raise standards of achievement (Black and Wiliam, 1998). Thus it is important for summative assessment to use methods that are complementary to, and do not compete with, formative assessment processes.

There is conflicting evidence as to the impact that a greater role for teachers in summative assessment can have on students. There is concern that the dual role of the teacher can affect student/teacher relationships. This could have a beneficial or detrimental impact, through the effort students put into their work or through the anxiety they have about being constantly assessed, with consequences for performance.

In relation to teachers, it is argued that, again, the impact may be positive or negative. On the one hand, it is suggested that a summative assessment role adds to teachers' workloads and does not produce outcomes in which users can have confidence. On the other hand, there are arguments that taking part in the processes required for summative assessment sharpens understanding of learning objectives, focuses teaching on the full range of outcomes, adds to professional competence and, allied with efficient moderation procedures, supports greater dependability in assessment. Moreover it is much less costly than external tests and examinations.

Similarly, there is some evidence that the teacher's role in summative assessment is associated with a broadening of the curriculum to encompass those outcomes that can be assessed by teachers but not by external tests. At the same time it is possible that the focus of teaching on what is assessed may have the reverse effect on an already broad curriculum.

These arguments are leading to suggestions that a greater role should be given in summative assessment to teachers' assessment for students. Thus a review of research on the impact that teachers' assessment can have on practice is relevant at this particular time, in order to identify what research can tell us about current and past practice in using assessment by teachers and so inform policy decisions about possible change.

Thus the aims of this review are to investigate the extent to which educational research provides evidence of the nature of the impact of teachers' summative assessment in these three areas: on students, on teachers and on the curriculum. Evidence on the particular circumstances of impact was sought so that, where trustworthy evidence was found, implications for policy and practice could be identified.

The outcomes were as follows:

- The production of a map of studies reporting on the impact of using teachers' assessment for summative purposes on students, teachers and the curriculum

- The identification (ideally through a process of consultation with users – although the reduced timescale for the review reduced the extent of this consultation before submission of the report) of the implications of the findings for different user groups, including practitioners, policy-makers, those involved in teacher education and professional development, employers, parents and students

- Publication of the full report and of short summaries for different user groups in the Research Evidence and Education Library (REEL)
- Identification of further research that is needed in this area.

1.6.2 Review questions

The considerations reported in this section led to the identification of the main review question as:

What is the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes?

To achieve its aims the review addressed the subsidiary question:

What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes?

The findings are used to address the further question:

What are the implications of the findings for policy and practice in summative assessment?

1.6.3 Scope of the review

The review has considered evidence from studies of teachers' assessment used in the context of summative assessment in schools, with students aged from 4 to 18. It sought studies of assessment by teachers in all curriculum areas and for both 'internal purposes' (reporting within the school and to parents) and 'external purposes' (where outcomes are used for certification and for accountability). It includes studies of assessments conducted using evidence from observation during regular activities or the use of classwork (or coursework) assessed against common criteria, and those where the assessment is based on special tasks or projects assessed by the teacher.

It was anticipated that there would be some studies in which the data reported were qualitative and may take the form of case studies, while others would be likely to report statistical or judgemental evidence of impact made by teachers. It was also intended that all such studies meeting the inclusion criteria would be included in a map of the types, designs and topic focus of studies, created as part of the review.

2. METHODS USED IN THE REVIEW

2.1 User involvement

This chapter describes how the review was carried out. The account of user involvement is followed by an outline of how the EPPI-Centre review procedures were implemented. These included procedures for searching for and documenting studies; applying inclusion and exclusion criteria; keywording; mapping included studies in terms of the keywords; in-depth data extraction; and synthesis of findings. It ends with information about quality-assurance procedures.

2.1.1 Approach and rationale

The users of this review include all involved with education. However, the review is concerned with matters relating to the impact of assessment by teachers on students, teachers and the curriculum, which are of particular interest to those making decisions about policy at national, local and school level. Thus the main focus is to inform policy-makers concerned with assessment and practitioners and their professional bodies. The direct involvement of users in the conduct of the review is through membership of the Review Group. The Assessment and Learning Research Synthesis Group (ALRSG) includes the following users: a secondary school deputy head teacher with responsibility for assessment, a local authority primary adviser, a project director of the National College of School Leadership. Two members of the group are members of AAIA, another led the review of assessment in Wales and another is Director of the Learning to Learn project of the ESRC's Teaching and Learning Research Programme. Eight of the Review Group are members of the Assessment Reform Group and through this the Review Group has an ongoing relationship with the DfES and with the Qualifications and Curriculum Authority (QCA).

2.1.2 Methods used

Users have been involved in the review in four ways:

- As members of the review group, in attending regular meetings to advise at key points of the review and at other times through email. Three meetings were held during the first six months of 2004.
- Providing information about studies through personal contact.
- Participating in keywording and in data extraction (four members were actively involved in this way).

2.2 Identifying and describing studies

2.2.1 Defining relevant studies: inclusion and exclusion criteria

Inclusion criteria

The search for and selection of studies was guided by the following inclusion criteria:

- *Language of the report*: Studies included were written in English. Databases and journals primarily in languages other than English were not searched.

- *Types of assessment*: Studies were included which dealt with some form of summative assessment conducted by teachers. Studies reporting on purely formative assessment by teachers were not included, but those where the assessment was used for both formative and summative purposes were included.

- *Study population and setting*: Studies were included which dealt with assessment procedures and instruments used by teachers for assessing students, aged 4 to 18, in school.

- *Study type and study design:* Studies were included if they reported information about changes in students, teachers or the curriculum that could be ascribed to the process of assessment by teachers being used for summative assessment. Both evaluations of naturally occurring and researcher-manipulated interventions were considered to be relevant. Surveys of students' and teachers' perceptions of the impact of using assessment by teachers for summative purposes and case studies of situations in which teachers' assessment is used for these purposes were also included.

- *Topic focus*: Since teachers' assessment can be used in all subjects, studies from any curriculum area were included. These included assessment where the task is decided by teachers and the outcome judged against common criteria, and where teachers use tasks and criteria prepared by others. They also included studies of assessment used of internal school purposes and those used of external purposes, such as student certification or school accountability.

The full set of inclusion and exclusion criteria used to define the study is given in Appendix 2.1 and section 3.1 gives the results of applying the inclusion and exclusion criteria.

2.2.2 Identification of potential studies: search strategy

Studies were identified through a combination of (i) a two-stage strategy, used for databases and citations in already identified reports, where there is no immediate screening, and (ii) a one-stage strategy, where handsearching, either by hand or online, allowed immediate screening.

The two-stage search was begun by searching bibliographic databases (ERIC, BEI) and references in key publications. Details of the search strategies for electronic database are given in Appendix 2.2.

The one-stage search was begun by creating a list of relevant journals from references in key studies already obtained and building on previous reviews. Those journals online were searched by computer; other journals held in the library were searched by hand, as were back numbers of those only recently put online. Details of journals handsearched are given in Appendix 2.3. Titles and abstracts were reviewed in relation to the inclusion and exclusion criteria before being entered into the database. Other papers were found by searching specialist websites (NFER, ERSDAT, CRESST), by making requests to members of relevant associations, other review groups, and using personal contacts. All papers identified in these ways were included in an EndNote database, each being labelled with its source and method of identification.

2.2.3 Screening studies: applying inclusion and exclusion criteria

Screening of titles and abstracts entered into the database was carried out by the author in order to check that they all met the inclusion and exclusion criteria. Each excluded paper was labelled with the reason(s) for exclusion. Those papers judged as meeting the inclusion criteria were entered into a second database. Full reports were obtained, where possible, for these papers and the inclusion criteria re-applied to the full text; those not meeting these criteria were excluded and labelled according to the reason.

Reasons for exclusion were recorded for each paper, as follows:

A: Not summative assessment. Studies were excluded if information was gathered for formative purposes only; also excluded were aptitude tests and special needs assessments.

B: Not assessment by teachers. Studies were excluded if they reported assessment of teachers or studies of school evaluation; also excluded were studies of teacher administered tasks or portfolios that were graded externally.

C: Not related to education in school. This excluded studies relating to college students; higher education; nursing education or other vocational education.

D: Not reporting impact of the process of assessment on students, teachers or the curriculum. Studies were excluded if the impact reported was a result of the outcome of the assessment and not the process.

E: Not research. Studies were excluded if they did not report empirical study of particular procedures of assessment by teachers; also excluded were handbooks and reviews, and reports of instrument development or description, without a report of their use.

2.2.4 Characterising included studies

The included studies were keyworded using EPPI-Centre Core Keywording Strategy (EPPI-Centre, 2002a). Additional keywords specific to the context of the review, with guidelines for application, were added to those of the EPPI-Centre. The EPPI-Centre Keywords and the review-specific keywords are given in Appendix 2.4.

Keywording of all included studies for which it was possible to obtain full texts was carried out by two people working independently. The author keyworded all the studies. The second keyworder was either a research assistant or a member of the Review Group.

Keywording resulted in the exclusion of some studies that were found not to meet the inclusion criteria. The agreed keywords for the remaining studies were used to produce the systematic map of included studies. All the keyworded studies have been added to the larger EPPI-Centre database, REEL, for others to access via the website.

2.2.5 Identifying and describing studies: quality-assurance process

Records were made of all searches: electronic database searches were documented; dates of journals searched were recorded. The author's judgements about inclusion and exclusion criteria were checked by EPPI-Centre staff for a sample of the papers (33 of 343 studies). All studies were keyworded by two people and any differences were resolved by discussion. Staff of the EPPI-Centre also carried out a quality-assurance role in applying inclusion and exclusion criteria and in keywording a sample of studies (9 of 26 studies).

2.3 In-depth review

2.3.1 Moving from broad characterisation (mapping) to in-depth review

The studies were 'mapped' in terms of the keywords and various tables presented to a meeting of the Review Group. It was decided that all 23 studies remaining after keywording appeared to be equally relevant to the review questions and should be included in the in-depth data extraction.

2.3.2 Detailed description of studies in the in-depth review

The 23 keyworded studies were entered into the EPPI-Centre's detailed data extraction software, EPPI Reviewer and data extracted using EPPI-Centre generic (EPPI-Centre, 2002b) and review-specific questions relating to the weight of evidence of each study in the context of the review.

2.3.3 Assessing quality of studies and weight of evidence for the review question

In order to ensure that conclusions were based of the most sound and relevant evidence, judgements were made using the EPPI-Centre ‘weight of evidence’ criteria. This involves judgements about three aspects of each study (A, B, C) and the combination of these to give an overall judgement of the weight that could be attached to the evidence from a particular study to answer the review question (D).

The criteria for assessing weight are as follows:

A: Soundness of methodology

Judgement of how well the study had been carried out was informed by the responses to questions about the internal methodological coherence during the data extraction. These answers were given on the basis of the information in the study report, which may or may not have given an account of all aspects of the study required for judging the soundness of the research. The judgement of methodological soundness was thus dependent on what was reported in the study by the authors. The lack of information about a certain feature did not necessarily mean that this feature was not attended to in practice by the study. Studies were rated as high, medium or low in relation to methodological soundness, according to what was reported. This judgement was not review-specific.

B: Appropriateness of research design for answering the review questions

The second judgement was made in relation to the extent to which the type and design of study enabled it to be used to address the review questions. In theory, some study types or designs might be better matched than others to the focus of the review. This was not a judgement of the value of the study in its own right, but only in respect of how well its design enabled the review questions to be answered and was thus review-specific. Studies were rated high, medium and low in relation to this aspect.

C : Relevance of the particular focus of the study for answering the review questions

As in B, this judgement concerns the match of the study to the purposes of the review and is not a judgement on the value of the study *per se*. In this case, the aspect of interest is the topic focus (including conceptual focus, context, sample and measures) of the study: that is, how well the nature of the data collected helped to answer the review questions. Again, the judgements were review-specific and made in terms of high, medium or low relevance.

D: Overall weight that can be given to the evidence in relation to the review focus

The judgements for the three aspects were combined into an overall weight of evidence towards answering the review question. In doing this, where there was a difference of judgement between A, B and C, the overall judgement was based on the majority rating but with the condition that the overall weight could not be higher than the weight for C. The rationale for this was that a study judged to be giving evidence of only medium weight on account of relevance of focus, context, sample and measures could not provide high weight of evidence overall from the review.

2.3.4 Synthesis of evidence

The structure for the synthesis of evidence from the in-depth review was taken from the review question: What is the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes? There were potentially three main headings, bringing together the findings for impact on students, teachers and the curriculum. However, since few studies were found that dealt with the curriculum – and those that did were also concerned with impact on teachers – these were combined. Within these broad groups formed according to the nature of the impact reported, there are important subdivisions according to whether the assessment is used for internal school purposes only (such as grades and routine school tests and examinations) or for use by others outside the school (as in the case of certification, selection, transfer, or the accountability of the school). Although it was often the case that an assessment would be used for both internal and external purposes, it was thought helpful to try to identify the main use and to discuss studies under these subheadings.

Thus the first two sections of the synthesis are concerned with studies giving evidence mainly on impact on students and impact on teachers, teaching and the curriculum. Evidence relating to the subsidiary question – What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes? – is discussed in a third section of the synthesis under five headings that were identified as the main conditions affecting the impact.

2.3.5 In-depth review: quality-assurance process

All in-depth data extraction was carried out independently by at least two people, using the generic data extraction guidelines on EPPI-Reviewer and review-specific questions. The author, three research assistants and four members of the Review Group between them conducted 46 data extractions. For each study, those completing independent data extractions compared their decisions and came to a consensus by direct communication. Five studies were also data extracted by a member of the EPPI-Centre staff for quality-assurance purposes and again any differences were resolved by discussion.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

This chapter presents results of the stages of searching, screening using inclusion and exclusion criteria, and the application of the generic EPPI-Centre and review-specific keywords. The numbers of studies at the various stages of filtering studies are given in a flow diagram. The characterisation of the selected studies in terms of the keywords is described and the results are given of the quality-assurance procedures for these parts of the process.

3.1 Studies included from searching and screening

The number of papers and studies at different points in the searching and screening processes are summarised in Figure 3.1 It can be seen that the total number of papers screened was 343.

Table 3.1 indicates the source of the initial papers found and, for comparison, the source of the studies that were included in data extraction.

Table 3.1 Results of initial search (343 papers)

Identification	Number (%)	Number included (%)
Two-stage screening		
ERIC	231 (67)	1 (4)
BEI	10 (3)	0
Citation	29 (8)	5 (22)
One-stage screening		
Electronic database (ERSDAT, NFER, CRESST)	5 (1)	2 (9)
Handsearch (not JOL)	51 (15)	13 (57)
Journal online (JOL)	6 (2)	2 (9)
Contact	11 (3)	0
Total	***343***	***23***

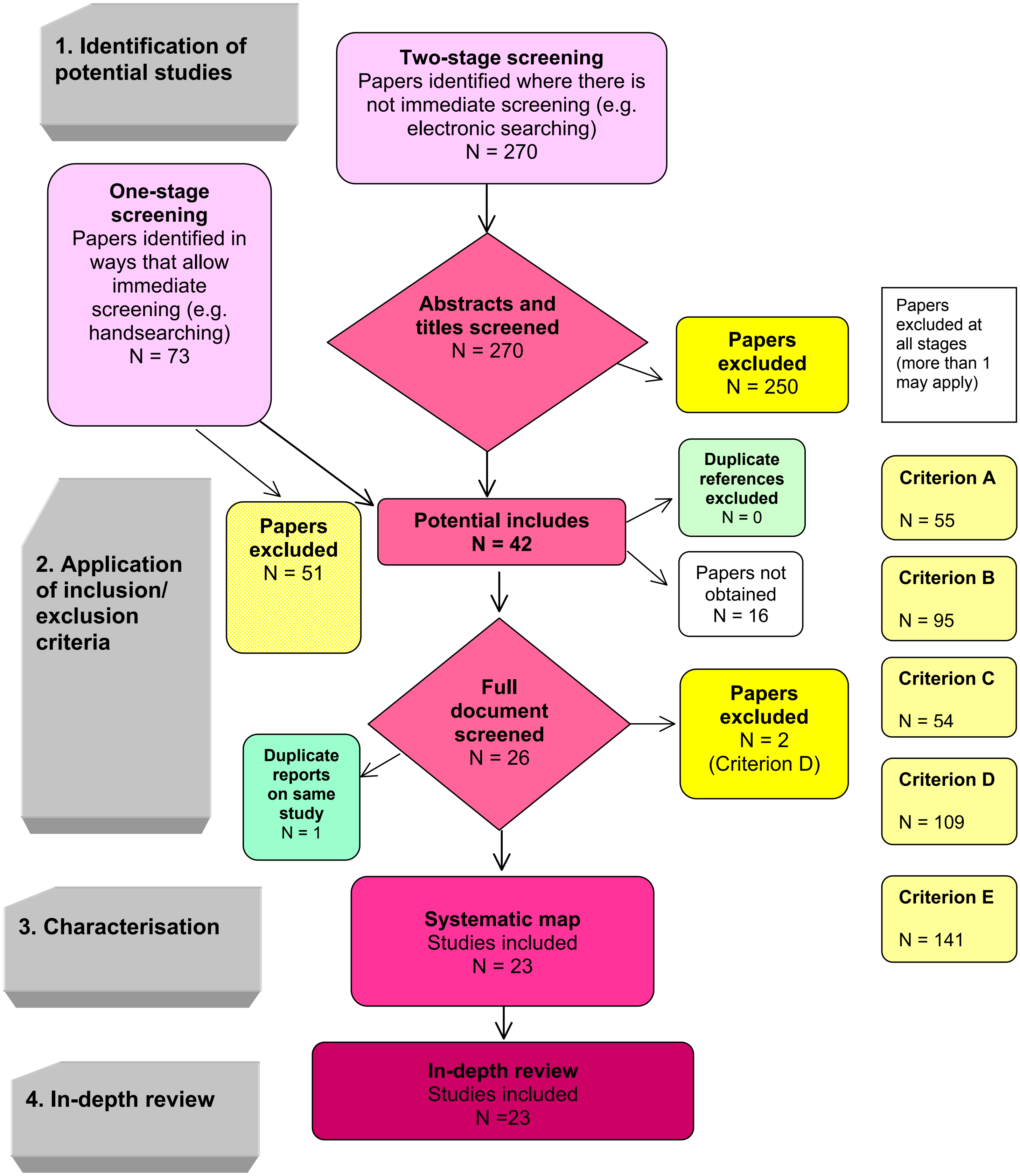
The criteria for excluding papers and the number excluded at all stages are given in Table 3.2. There were 303 papers excluded, some being excluded for more than one reason, whilst 16 others were unobtainable. Two papers were linked, and one was then excluded from the data extraction as a separate item so that only one of these appears in the list of studies used in data extraction.

Table 3.2 Exclusion criteria and numbers excluded at all stages (not mutually exclusive)

	Criteria (more than one can apply)	Number of studies
Criterion A	Not summative assessment. Studies were excluded if information was gathered for formative purposes only; aptitude tests and special needs assessments were also excluded.	55
Criterion B	Not assessment by teachers. Studies were excluded if they reported assessment of teachers or studies of school evaluation; also excluded were studies of teacher administered tasks or portfolios that were graded externally.	95
Criterion C	Not related to education in school. This excluded studies relating to college students; higher education; nursing education or other vocational education).	54
Criterion D	Not reporting impact of the process of assessment on students, teachers or the curriculum. Studies were excluded if the impact reported was a result of the outcome of the assessment and not the process.	109
Criterion E	Not research. Studies were excluded if they did not report empirical study of particular procedures of assessment by teachers; also excluded were handbooks and reviews and reports of instrument development or description, without a report of their use.	141

In the screening process all papers were labelled either IN or OUT with the reasons for exclusion. In addition, some papers, considered to be of particular relevance but excluded for one of these reasons, were labelled as useful for the background discussion. Of the 42 papers labelled IN, the full texts of 16 could not be found, leaving 26 for the keywording stage. At this stage two further papers were excluded, using the same criteria as above. In addition since the same data were used in two papers these were linked, and one was subsequently excluded as a separate item with only one of them included in the data extraction. Thus 23 studies remained in the systematic map and in-depth review.

Figure 3.1: Filtering of papers from searching to map to synthesis



3.2 Characteristics of the included studies

3.2.1. Characterisation in terms of the EPPI-Centre keywords

The classification of the 23 included studies in terms of the keywords is given in Appendix 3.1. Tables A3.1.1 to A3.1.5 give the classification according to the EPPI-Centre keywords. These show that just over half of the studies (12) were conducted in parts of the UK mainly England, nine in the United States, one in New Zealand and one in Hong Kong. All the studies were characterised as focusing on learners in recognition that the assessment involved learners. However, in 19 studies, it was reactions, actions and views of teachers that were the main object of interest. The majority of studies concerned students in primary and secondary schools between the ages of 5 and 16 years and all dealt with students of both genders. Four studies were classified as ‘description’ four as ‘exploration of relationships’, 12 as evaluations of naturally occurring interventions and three as evaluations of interventions introduced by the researchers.

3.2.2. Characterisation in terms of the review-specific keywords

Tables A3.1.6 to A3.1.10 in Appendix 3.1 give frequencies relating to classification using review-specific keywords. Thirteen studies give some report of impact of the teachers’ assessment on students and 17 reported impact on teachers or teaching. Only two studies report impact on the curriculum and as these also report impact on teaching, they were combined into one group in reporting the synthesis of studies. The assessment was mainly concerned with the subjects that form the core of the curriculum in England: English, mathematics and science. One study referred to grading across the curriculum and thus included arts subjects, but assessment in these subjects was not specifically reported upon. Other subjects involved include geography and technology. Sixteen studies concerned summative assessment tasks set by the teacher, while ten concerned teachers using externally prescribed tasks, in some cases selected from a range of such tasks.

Fifteen studies report on assessment used for internal purposes, producing grades or levels that were used for school records, reports to parents etc. Generally these were ‘low stakes’ uses, since they were not used for certification of students or school accountability measures. These include studies of teachers’ assessment in the National Curriculum Assessment since only the standards task or test results were used for league tables and not the teachers’ assessment. However, internal school use may have increased the stakes for teachers in some schools. Teachers’ assessment contributed to students’ results with explicit high stakes use are reported in seven studies. In three studies, one of the uses of the assessment was for research purposes.

Fifteen studies concerned assessment tasks that could be described as part of regular work, which included work collected in portfolios. Eight studies concerned special activities, some of which were conducting projects.

Tables A3.1.11 and A3.1.12 show the use of different types and origin of assessment tasks represented in the studies. In Table A3.3.11 it is evident that projects were used for external purposes, while embedded tasks and regular work were used mainly for internal purposes. There were no studies reporting the use of embedded tasks for external accountability purposes. This is reflected in Table A3.3.12 where the majority of studies involved tasks set by the teacher used for internal purposes.

Tables A3.1.13 to A3.1.17 show how the impact reported in the studies varied with the origin of the task, the type of task, the use of the results, the educational setting of the study and the achievement assessed. There were only two studies reporting impact on the curriculum and in both these studies the tasks were selected by the teacher. There were more studies reporting impact when the tasks were set by the teacher than when they were selected or externally prescribed. However, these tables are not easy to interpret, mainly because the impact of the assessment was not the main focus of the studies, but was incidental to it. The studies were not in general designed to report on the impact of the assessment; nevertheless, they provide information that could be extracted for the purposes of this review.

3.3 Identifying and describing studies: quality assurance results

3.3.1 Applying inclusion and exclusion criteria

The application of inclusion and exclusion criteria to titles and abstracts was checked by EPPI-Centre staff for 33 of the 343 studies (approximately 10%). There was disagreement in four cases, which led to clarification of the criteria.

3.3.2 Keywording

For keywording, where all 26 studies were classified by two people, complete agreement was found for 18 studies. Differences were found equally in the EPPI-Centre keywords and the review-specific keywords. Differences in relation to study type occurred in the case of some studies where an evaluation of a naturally occurring intervention could equally be seen as ‘description’ and in some cases, ‘exploration of relationships’. Quality assurance by EPPI-Centre was carried out for nine studies. For these studies, difference between the two keyworders were discussed, which led to further clarification of the inclusion criteria as well as the interpretation of keywords and resulted in the exclusion of two studies. In one case, this decision was made after communication with the study author to clarify the procedures for the assessment. This revealed that the assessment tasks were externally marked and thus the study was excluded by criterion B.

4. IN-DEPTH REVIEW: RESULTS

This chapter describes the characteristics and findings of the 23 studies selected for in-depth review. The synthesis of findings in relation to the main review question is given in two main sections: those studies reporting impact of the process of teachers' summative assessment on students; and those where the impact reported is on teachers or teaching or the curriculum. Each of these sections is sub-divided according to the use of the results for internal or for external purposes. There is a summary of main points at the end of each sub-section. Findings in relation to the subsidiary question are reported under five headings, relating to the conditions which were identified as affecting the nature and degree of the impact of the process of teachers' assessment on students and teachers, with a concluding summary of main points.

4.1 Further details of studies included in the in-depth review

An outline of the aims, study type, data collection, data analysis, findings and conclusions of the 23 studies from which data were extracted is given in Appendix 4.1. Table 4.1 summarises information about the main focus of the studies, the age of the learners assessed and the judgements of weight of evidence from each study. As noted earlier (section 2.3.3), the judgement combining the three aspects A, B, and C into an overall weight of evidence for answering the review question was based on the majority rating but with the condition that the overall weight could not be higher than the weight for C.

Table 4.1: Classification of studies by object of the impact reported and weight of evidence

Study	Object of impact reported	Age of students assessed (years)	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Abbott *et al.* (1994)	Students Teachers/teaching	5-10	Medium	Medium	Medium	Medium
Bennett *et al.* (1993)	Students	5-10	High	Medium	Medium	Medium
Bennett *et al.* (1992)	Students Teachers/teaching	5-10	Medium	Medium	High	Medium
Brookhart and DeVoge (1999)	Students	5-10	High	High	High	High
Bullock *et al.* (2002)	Students	11-16	Medium	High	High	High
Carter (1997/8)	Students	17-20	Medium	Low	Low	Low
Cizek *et al.* (1995/6)	Students Teachers/teaching	5-10 11-16 17-20	Medium	Medium	Medium	Medium

Flexer *et al.* (1995)	Students Teachers/teaching	5-10	High	High	High	High
Gipps and Clarke (1998)	Teachers/teaching	5-10 11-16	High	High	Medium	Medium
Hall and Harding (2002)	Teachers/teaching	5-10	Medium	High	Medium	Medium
Hall *et al.* (1997)	Students Teachers/teaching	5-10	Medium	Medium	High	Medium
Hiebert and Davinroy (1993)	Teachers/teaching	5-10	Medium	Low	Low	Low
Hill (2002)	Teachers/teaching	5-10	Medium	Medium	Medium	Medium
Iredale (1990)	Students	11-16	High	High	High	High
Johnston *et al.* (1993)	Teachers/teaching	5-10 11-16 17-20	Medium	High	High	High
Koretz *et al.* (1994)	The curriculum Teachers/teaching	5-10 11-16	Medium	Medium	Medium	Medium
McCallum *et al.* (1993)	Teachers/teaching	5-10	Medium	Medium	Medium	Medium
Morgan (1996)	Teachers/teaching	11-16	Low	Low	Medium	Low
Pilcher (1994)	Students	11-16	High	High	High	High
Stables (1992)	Students Teachers/teaching	11-16	Medium	Medium	Low	Low
Valencia and Au (1997)	The curriculum Teachers/teaching	5-10 11-16	High	Medium	Medium	Medium
Whetton *et al.* (1991)	Teachers/teaching	5-10	Medium	Medium	Medium	Medium
Yung (2002)	Students Teachers/teaching	17-20	High	High	High	High

4.2 Synthesis of evidence: overall review question

What is the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes?

4.2.1 Evidence from studies reporting impact on students of teachers' summative assessment practices

Twelve studies provided evidence of some impact on students; six giving evidence of high weight, four evidence of medium weight and two of low weight, as shown in Table 4.2. Three of the high weight studies concerned assessment used for external purposes (potentially of high stakes for the student or school or both) and, in four cases, they concerned assessment for internal school purposes (for grading, school records and reporting to parents); in one study, the assessment was intended to serve both internal and external purposes. In this section, these purposes are used as a structure for synthesising findings, reporting mainly the findings from high weight studies with reference made to evidence of less weight where relevant.

Table 4.2: Studies providing information on the impact of teachers' summative assessment on students

Study	Age of students assessed (years)	Type of study	Use(s) of result	Overall weight of evidence
Brookhart and DeVoge (1999)	5-10	Exploration of relationships	Research – theory building	High
Bullock *et al.* (2002)	11-16	Evaluation: naturally occurring	External for certification (high stakes for student)	High
Flexer *et al.* (1995)	5-10	Evaluation: researcher-manipulated	Formative and summative Internal (for grading, in-school records, reporting to parents) Research	High
Yung (2002)	17-20	Description	External for certification (high stakes for student)	High
Iredale (1990)	11-16	Evaluation: naturally occurring	Internal (for grading, in-school records, reporting to parents) External for certification	High
Pilcher (1994)	11-16	Description	Internal (for grading, in-school records, reporting to parents)	High
Abbott *et al.* (1994)	5-10	Evaluation: naturally occurring	External for accountability	Medium
Bennett *et al.* (1993)	5-10	Exploration of relationships	Internal (for grading, in-school records, reporting to parents)	Medium
Cizek *et al.* (1995/6)	5-10 11-16 17-20	Exploration of relationships	Internal (for grading, in-school records, reporting to parents)	Medium
Hall *et al.* (1997)	5-10	Evaluation: naturally occurring	Internal (for grading, in-school records, reporting to parents)	Medium
Carter (1997/8)	17-20	Evaluation: researcher-manipulated	Internal (for grading, in-school records, reporting to parents)	Low
Stables (1992)	11-16	Evaluation: naturally occurring	External for accountability	Low

Studies where the assessment is for external purposes of certification and/or accountability

Bullock *et al.* (2002) studied the effects of teacher-assessed coursework which contributed to the award of the General Certificate of Secondary Education in England for students at age 16. The coursework was intended to 'raise the validity of the assessment process and enhance the learning of students' (Bullock *et al.* 2002, p 326). The subjects chosen for study were geography, for which the coursework was one or two substantial pieces of fieldwork involving gathering, analysing and reporting data from an out of school location; and English, for which the coursework was a portfolio of essays and records of oral presentations. The researchers conducted a series of interviews with students (on two occasions) and with their teacher and parents (on one occasion) to gather data about perceptions of the impact of coursework on creative and critical thinking and independent learning.

The researchers found a positive impact of the coursework on students' learning:

> …all the pupils in the sample claimed some degree of independent learning as a result of their course work. Those with a positive attitude to school (the great majority) found the course work motivating on account of the different skills that are practised and assessed through course work tasks. There were perceptions of differences in the skills promoted by the two subjects areas. It was claimed that geography course work tended to encourage literacy, numeracy, teamwork skills, and an ability to use initiative. On the other hand, it was suggested that English course work encouraged insight, originality and imagination (Bullock *et al*., 2002, p 329).

The students perceived the coursework as motivating, helping them to retain knowledge and skills, and helping them to find things out for themselves. Reasons for this given by the students, teachers and parents were in terms of the nature of the coursework, which provided some element of choice, and the use of different techniques to find things out or express themselves, and because the work was assessed.

> On balance pupils enjoyed course work. Together with their parents and teachers, they valued the distinctiveness of this mode of learning and nature of the skills they acquired from it. ...In general, students were receptive to opportunities to learn more about learning – organising and managing their learning, how to become more independent – and demonstrated this in the interviews. Further benefits claimed were ownership of the project and freedom to organise their own learning and modes of working. The downside was the demands of time, deadlines, and the pressures from concurrent, similar work in other subjects (Bullock *et al*., 2002, p 330).

The researchers, however, found teachers were less convinced of the effect on student learning than the students and their parents. They were aware of the constraints of the assessment criteria and were often reluctant to allow students to take control of the coursework. There was also evidence that teachers had not communicated assessment criteria effectively to students.

> Teachers assume that students will perceive the demands of learning and assessment in the same way that they do. In fact, despite teachers'

> assertions that marking schemes have been shared with students, the students tend not to understand what the assessment criteria actually require from them. Our research suggests that it is not sufficient to tell them; illustrations, examples and models are required (Bullock *et al*., 2002, p 338).

There was also awareness on the part of both students and teachers that what gained credit was the *product*, not the *process* by which it was achieved. This left open the potential for giving emphasis to the product by teaching specific techniques needed to reach higher-level grades, which most teachers were reported as doing. Thus the pressure on teachers for students to achieve higher grades meant that the completion of coursework had less value for developing students' independent learning.

An innovative approach to assessment in science in the secondary school in England was studied by Iredale (1990). The Graded Assessment in Science Project (GASP) was developed in the 1980s in order to replace the single examination at the age of 16 with a series of school-awarded certificates accumulated throughout the five years of secondary education. Each certificate assessed some content (by testing, using items drawn from a bank), process skills (assessed by the teacher in a variety of contexts) and explorations (drawn from a range of tasks and incorporated into classwork). Iredale studied the attitudes to the GASP scheme of first and second year secondary students, using data collected by a questionnaire (completed anonymously), interviews (two pupils at a time) and students' essays. She reports a general impression from the results of 'enthusiasm and approval' for the scheme. Responses for girls and boys were very similar.

> A large majority of pupils (83%) believe that passing levels will help them in the future. At this early stage in their secondary school career some pupils appear to be looking ahead to GCSE and beyond. They see the value of the GASP scheme in providing a continuous-assessment route to GCSE and evidence of their achievements in science. …91% of pupils considered the scheme preferable to an end of year examination (Iredale, 1990, p 134).

In their essays and interviews, students commented with approval that not everything depended on what was done at the end of a year or course and that it was fairer 'on those people who get worried when there's a big examination' (Iredale, 1990, p 134). There was, however, a minority of 16% who felt pressure due to the high frequency of testing (on average one test every four weeks). Iredale inferred from the data that these were the students who found the tests too difficult. This led her to suggest that the effect 'on these pupils of the constant reminders of their failure to progress which the GASP scheme gives, needs to be considered' (Iredale, 1990, p 137).

In the light of the findings of Bullock *et al.* (2002) on the need for better communication of goals and criteria to students, it is significant that Iredale found that only about one-third of students felt that they understood the GASP scheme well. She concluded that 'there seems to be a case for making the scheme easier for pupils to understand. Descriptions and examples of skills, for example, may be preferable to code letters and labels' (Bullock *et al.,* 2002).

This criticism of teacher-based summative assessment in high weight studies found support in the low weight evidence from Stables (1992). This was a cross-sectional study of how teachers in four classes went about assessing speaking and listening. Data were collected by observation of lessons and interviews with teachers and discussions at periodic network meetings. The assessment of process in oral work was a focus of discussion among teachers, who felt that 'too little is known about the processes of oral work' (Stables, 1992, p 111) so that teachers were questioning what they were looking for. Not surprisingly, then, the students were also in the dark:

> An important consideration, if process is to be assessed in oral work, is that the need for explicit discussion, planning, etc., is made clear to the pupils involved, and that work is done specifically on improving those skills. The observed lessons all pinpointed situations which can stimulate such discussion, but it is unclear whether a) stimulation of discussion is enough, and b) whether pupils appreciate the importance of the fullest possible development of the discussion. (Stables, 1992, pp 111-112)

The focus of the study by Yung (2002) was, as in the case of Iredale's study, the replacement of an external examination by teacher-based assessment. In Hong Kong, at some time in the early 1990s, the Advanced Level biology practical examination was replaced by the Teacher Assessment Scheme (TAS). It was expected and intended that this would have 'a liberating influence on the curriculum and thus bring advantages to teaching and the students'. The study, providing evidence of high weight for the review, was designed to illuminate what was happening in the classroom when the TAS was being used to assess students' practical work. Yung reported, from data collected in a wider study, case studies from classroom observations and interviews with three teachers whose practices in implementing the innovation varied.

The findings were presented as accounts of the events from the teachers' point of view. Although there was no direct evidence of the impact on students, the different practices were clearly associated with different assessment conditions for the students. Case 1 was a teacher (Ivor) who kept very closely to the TAS regulations. He followed the suggestions for using the assessment formatively, both in his teaching and in discussions later with students. He explained this in terms of making him a better teacher and the interpretation was that he did this in his own interests rather than in the students'. His approach to the regulations was 'readerly'. (This is a reference to the distinction originating in the work of Barthes (1975) between 'readerly' texts, that leave little room for interpretation and 'writerly' texts when the reader has an interpretative role; the terms can be used to describe the approach of the reader as well as the nature of the text.) Ivor kept very close to the regulations and described his role as 'like a policeman', finding fault, in order not to be accused of cheating.

> For Ivor, despite his stated desire to improve both his teaching and his students' learning...the introduction of the TAS with its many regulations was seen as imposing severe constraints upon his professional autonomy. Under such circumstances teacher professionalism was severely compromised as Ivor struggled to make sense of his changing role and responsibilities both as an assessor and as a teacher. Most importantly, underpinning his 'readerly' interpretation of the TAS regulations was his consciousness of 'protecting himself' rather than defending his students' interests. (Yung, 2002, p 104)

Case 2 was a teacher (Carl), who took a 'writerly' approach to the regulations. 'With a high level of professional confidence, Carl was able to impose his own professional interpretations on the TAS regulations, to balance its demands against other professional priorities and to exploit to the maximum what remained of his professional autonomy' (Yung, 2002, p 105). Thus he allowed students to work in ways that were more natural (in groups, conferring with each other and looking up references). 'The most significant influence on Carl's teaching was not the formal apparatus of external obligation and controls imposed on him by the TAS regulations, but his personal sense of professional obligation to offer students an all-round education' (Yung, 2002, p 107).

Case study 3 was a teacher (Eddy) who also did not keep exactly to the regulations, but with a different motivation – of making sure that students got high grades. He was exam-oriented in focusing just on what would help students to get high marks. 'Evidence suggests that Eddy's consciousness was directed towards his own self-interest rather than his students' learning. This in turn led to his own low level of professional confidence and, hence to a seemingly writerly but actually 'readerly' interpretation of the regulations.' (Yung, 2002, p 111)

From these cases, and from a wider study from which these cases were extracted, the author drew conclusions consistent with literature that teachers who adopt a critical stance to policy change are able to exercise control on own teaching. Variations in how teachers adopt this position is a function of their own professional confidence, which is guided by professional consciousness. The impact on students of the TAS depended on the professional consciousness of the teacher. Teachers who protect their own interests and keep rigidly to regulations find a conflict between being a teacher and an assessor. Teachers who see fairness to students and 'an all-round education' as priority will apply regulations flexibly to ensure this. Other teachers may use the regulations flexibly to ensure that students get high grades.

The extent to which teachers can adopt a different role in assessing arose in the study of the assessment of young students by Abbott *et al.* (1994). This study, providing evidence of medium weight also concerned assessment by teachers, using an externally prescribed task and marking scheme. The context was the administration of the Standard Assessment Tasks (SAT) for 7-year-olds as part of the National Curriculum tests in England in the early 1990s. Administration tests were observed in classes in three schools, followed by interviews with teachers after the administration. A considerable variation in practice was observed, in, for example, the time given to introduce the task, the time for students to complete the task (there was intended to be no time limit) and interpretation of the marking criteria. The individual administration imposed unusual restraints on the students, who were used to working in groups and found the individual situation unusual and may not have done their best work. Teachers were uncomfortable about not being able to respond to other children when they were assessing one child. The teachers also noted their 'dislike of asking children to work alone, of being unable to offer help if asked without affecting the levels recorded' (Abbott *et al*., 1994, p 167). At the same time, they were worried about the subjective judgements involved in SAT procedures 'so it is arguable that teacher assessment is as trustworthy as SAT testing over most areas and can fulfil diagnostic and formative aims' (Abbot *et al*., 1994, p 171). The researchers questioned whether it is possible that assessment procedures for young students can be standardised in the same way as is possible for older students; they cast doubt on the reliability of the SATs for the purposes of external reporting.

Main points about impact on students of summative assessment for external purposes

There is high weight evidence of the following:

- Older students respond positively to summative assessment by teachers of their coursework, finding the work motivating and being able to learn during the assessment process (Bullock *et al*., 2002).

- Students need more help, in the form of better descriptions and examples, to understand the assessment criteria and what is expected of them in meeting these criteria (Bullock *et al*., 2002; Iredale, 1990; Stables, 1992).

- The impact of summative teacher assessment on students depends on the high stakes use of the results (Yung, 2002).

- The impact of summative teachers' assessment on students will be affected by the way teachers interpret their roles as assessors and by their orientation towards improving the quality of students' learning or maximising their marks. (Bullock *et al*., 2002; Yung, 2002).

There is medium weight evidence of the following:

- Teachers consider that young students may not do their best work when constrained by an external task (Abbott *et al*., 1994).

Studies where the assessment is for internal purposes: grading, school records or reporting to parents

There were two main findings of relevance to this review from the high weight evidence in a study by Brookhart and DeVoge (1999). These concerned the impact of feedback from earlier work on the self-efficacy of students to do current work and how the teacher presents and treats classroom assessment events. The study tested a theoretical model for interpreting results of assessment events in a limited environment. The authors' premise is that classroom achievement is conventionally measured by classroom assessments that teachers construct or select for this purpose. These assessments are the basis of students' perceptions as to what it is important to learn and where to direct effort in learning. To explore these relationships, two third Grade language arts classes were studied over four classroom assessment events. A description of the level of perceived task characteristics, perceived self-efficacy, amount of invested mental effort, achievement, and the relations among these was sought. Four different classroom assessment events were selected for study in each class, in consultation with the teachers. For each event, a pre-survey was administered to the whole class to collect perceptions of perceived task characteristics and perceived self-efficacy to do the task. A post-survey was administered after the assessment but before students received feedback, to collect perceptions of amount of invested mental effort.

In relation to feedback, it was found that students' self-efficacy judgements about their abilities to do particular classroom assessments were based on previous experiences with similar kinds of classroom assessments. Results of previous spelling tests, for example, were offered as evidence about how students expected to do on the current spelling test. This finding is consistent with the model tested and also with self-efficacy research. The nature of the feedback was an important factor in this effect. Judgemental feedback (in the form of marks or comments such as 'good') is likely to be used by students as evidence of their ability to succeed in further activities of this kind, and may reduce effort in further tasks of the same kind if the feedback is negative. On the other hand, non-judgemental feedback that gives information about how to do better next time is likely to promote effort on a future occasion.

In relation to the presentation of assessment tasks, Brookhart and De Voge (1999) concluded that teachers' explicit instruction and how they present and treat classroom assessment events affects the way students approach them:

> When a teacher explicitly exhorts students to work 'to get a good grade' that teacher is on the one hand motivating students and on the other setting up a performance orientation that ultimately may decrease motivation. Teachers should make a point to exhort students to work for the satisfaction of learning or for its usefulness in accomplishing future work. (Brookhart and DeVoge, 1999, pp 423-424)

The authors question whether this is possible as long as the classroom assessment environment is affected by school district report card policies and the requirement to give grades.

A study providing evidence of low weight for the review confirmed the value of non-judgemental feedback to students (Carter, 1997/8). Carter's study of her own work with gifted mathematics students involved students in analysing and correcting their own tests before being given a grade. The author claims that the opportunity to identify, correct and explain their errors resulted in students taking responsibility for their own learning, removed some of the pressures from the test and led to a decline in careless errors.

The high weight study by Flexer *et al.* (1995) was concerned with the effect on teachers and students of professional development aimed at changing the teachers' assessment practices. The research, involving 14 third Grade teachers in three schools (five in each of two schools and four in a third), was carried out in a school district with a standardised testing programme in place, where the district was willing to waive standardised tests for two years in the schools taking part. The research team collected data over one academic year (1992-93). The intervention took the form of a series of weekly workshops with teachers and researchers dealing with assessment in reading and mathematics, the focus alternating week by week. The data took the form of regular interviews with teachers and transcripts of the workshop sessions. Although the initial intention of the workshops was to help teachers expand their assessment repertoires, the teachers requested materials for teaching to match the new assessments and thus the scope of the workshops was extended to include teaching.

The researchers report that, at the start of the project, 'most teachers held fairly traditional views about what mathematics is important to teach, what instruction should look like, and how students should be assessed' (Flexer *et al*., 1995, p 8). By the end of the first year, they noted several effects on teachers and on students. Teachers were more often using hands-on activities, problem solving and asking students for explanations. They were also trying to use more systematic observations for assessment. All agreed that the students had learned more and that they knew more about what their students knew. These changes were not achieved without effort. All the teachers struggled with different teaching approaches and new assessments. Many felt overwhelmed but they received generally positive feedback from their own classes; the students had better conceptual understanding, could solve problems better and explain solutions. The teachers' response was to attempt further change in assessment and instruction practices and become more convinced of the benefit of such changes.

The issues surrounding grades and their impact on student learning were investigated by Pilcher (1994) in a study providing high weight evidence. She investigated the perceptions of the meaning given to grades by teachers, students and their parents. Pilcher (1994) conducted six case studies, each comprising the English and maths teachers, the student and one of the student's parents, gathering information by individual interview and from documentation. Documentation was used for 'verifying and guiding responses'. The focus was on the grades given every six weeks, incorporating the results of assignments and test given during that period. Summarising the results, Pilcher reports:

> This study suggests that grades represent a combination of achievement of course content, ability level of students, and effort applied in class. A high grade means the student had some combination of high test scores, high grades on writing assignments, and applied effort. A failing grade means total lack of effort. Even students who fail tests can usually receive a passing grade if they apply effort. (Pilcher, 1994, p 80)

She also notes that it is possible for students to have achieved more than is indicated by the grade or to be given a grade higher than their test scores would indicate. Although both mathematics and English teachers did not assign a score to attitude, 'teachers' comments about adjustments of grades for particular students indicated that teachers made inferences about attitudes when assigning grades' (Pilcher, 1994, p 82). She found that students were aware of how teachers were grading them and so interpreted their performance in similar ways to the way the grades were assigned. However, this was not the case for parents: 'parents perceived grades as reflecting their child's achievement level' (Pilcher, 1994, p 83).

Thus grades were being used as rewards and punishments, and not just to record the students' achievements. In other words, they were used as extrinsic motivation, incurring all the disadvantages for students' motivation for learning that this entails. Pilcher concludes that 'in their current and recommended states, grades are more harmful than beneficial to student learning. Using grades to control student behaviours does not teach students to value learning' (Pilcher, 1994, p 87). She suggests that educational reforms in assessment need to focus on practical methods for reporting as well as on alternative methods of assessment. This will involve questioning the purpose of grades and developing grading procedures that do not interfere with students' intrinsic value of learning. 'The positive and negative consequences students face from the value others attach to grades must be addressed.' (ibid).

In the study by Iredale (1990), outlined above, the achievement of each graded level was used for internal school purposes and for reporting to parents as well as leading to the external GCSE. While, as noted, the majority of students found the regular testing motivating through providing short-term achievable goals, by the same token some of those progressing only slowly through the levels felt discouraged by finding others several levels ahead of them. Indeed, Iredale found ‘a significant difference between levels of satisfaction expressed by those pupils on the higher levels and those on the lower levels. 44% of second years at level 1 or 2 expressed satisfaction compared with 93% at levels 5 and 6.’ (Iredale, 1990, p 136). Thus, although the meaning of the levels was more evidently based on achievement than in the case of the grades studied by Pilcher, nevertheless, the Graded Assessment Science Project (GASP) assessment was being used as extrinsic motivation.

Two studies providing medium weight evidence for this review were also concerned with the meaning of teachers’ grades and the information that influenced teachers in arriving at their judgements of students’ achievement. The study by Bennett *et al*. (1993) involved a total of 794 students in kindergarten (Kg) and Grades 1 and 2 in two school districts, one in Cleveland, Ohio, and one in the Bronx, New York, USA. The aim of the study was to test a model of the relationship between tested academic performance, behaviour, gender and teachers’ judgements of academic performance.

The study found that, while there were no gender differences in academic test scores and academic grades for Grades 1 and 2, girls were given consistently significantly higher behaviour grades than boys. For academic ratings, a gender differences (in favour of girls) was found only in Grade 1. In all instances, gender was significantly related to behaviour grade, with effect sizes ranging from 0.23 to 0.37. For the Kg students, behaviour grade consistently affected teachers’ academic judgements after controlling for gender, academic score and missing data. Effect sizes were large, at 0.34 in Cleveland and 0.49 in the Bronx. The authors report that their findings suggest that gender had a consistent effect on academic judgements that appears to have been mediated by teachers’ perceptions of behaviour and that this effect was slightly stronger in the first grade than in the second grade. They conclude:

> In all grades and in both districts, after controlling for tested academic skill and for gender, we found that teachers’ perceptions of students’ behaviour constituted a significant component of their academic judgements. In other words, students who were perceived as exhibiting bad behaviour were judged to be poorer academically than those who behaved satisfactorily, regardless of their scholastic skill and their gender. In Grades 1 and 2, however, boys were consistently seen as behaving less adequately than girls. As a result, teachers’ perceptions of boys’ academic skills were more negative than their perceptions of girls’ capabilities. (Bennett *et al.*, 1993, p 351)

The size of the differences was considerable and constant across school districts. In Grades 1 and 2, a change in behaviour grade produced only slightly less change in academic judgement than the same proportional change in academic skill. In Kg, the effect of change in academic skill was essentially the same as that for behaviour. Thus behaviour perception was found to be a potentially distorting influence on teachers’ judgement. The impact on the students was likely to be one of setting up a vicious circle, in which boys judged to be lower in achievement than is actually the case are given activities which do not stretch them and so can lead to boredom and further bad behaviour.

The authors note two implications of this:

> First, these data reinforce the need to supplement teachers' judgements with other objective evidence of academic performance when important decisions about students are made….The second implication is the need for more concerted effort toward making teachers aware of the potential influence of student behaviour on their academic appraisals. (Bennett *et al.*, 1993, p 353)

A survey of teachers' assessment practices was conducted by Cizek *et al.* (1995/6) using an opportunity sample of elementary and secondary school teachers at the beginning of a master's level course. The survey asked, *inter alia*, about the factors that the teachers considered when giving grades, the frequency of giving major and minor assignments, their knowledge of other teachers' grading practices and their knowledge of district policies relating to grading. The researchers found wide variation in assessment practices. For example, in relation to the information considered in assigning the grade for a particular task or test, between one-third and one-half of teachers took into account, variously, effort, student ability, the difficulty of the task, the level of performance of the class as a whole, as well as how correct the individual's answer was. When deciding a final grade at the end of a marking period, 61% took account of informal judgements of effort and conduct, and 52% of information about attendance and class participation. Over half of the teachers were not sure how their grading practice compared with other teachers in their school and 12% were not sure whether their district had a formal grading policy. Even in districts that had a policy (in the case of 52% of the teachers), a majority of teacher indicated that they were unaware or deliberately ignored those policies. Commenting, the researchers pointed out:

> The finding that teachers may be unaware of a macro-level policy may not be all that surprising, given ubiquitous bureaucratic inefficiencies in disseminating information. However, in this study, the teachers surveyed reported that they are generally unaware of their colleagues' practices. …several who acknowledged that they were unsure about what their colleagues did vis-à-vis assessment and grading also indicated that they preferred it that way. A recommendation that follows from these observations is that schools more actively pursue engendering cultures of collaborative reflective practice, especially related to assessment. (Cizek *et al*., 1995/6, p 175)

This study did not report on whether students and other users of grades were aware of how they were constituted. However, the variability is clearly a potential problem for students, especially in the light of feedback from one assessment affecting effort and self-efficacy in further tasks, as reported by Brookhart and DeVoge (1999).

The study by Hall *et al.* (1997) of teachers' assessment practices was conducted in a contrasting context to that of Cizek *et al.* (1995/6). It was conducted in 1993 at a time when teachers of 7-year-olds had for three years been required to assess students at the end of their second year of school against national criteria. Thus teachers had common criteria to use and, in many schools, this meant that teachers were using or developing whole school policies on assessment in line with the mandatory requirements for assessment. Teachers of 7-year-olds in 45 schools were interviewed, just after completing the record of their teachers' assessment. The findings showed that there was sufficient commonality in approaches for the researchers to identify a 'model' or series of stages through which teachers go in conducting their assessment. This begins with planning suitable tasks to be incorporated into regular work to assess specific skills and ends in the decision about a level for each child.

In relation to the impact on students, Hall *et al.* (1997) report that 63% of teachers perceived the impact on students as positive. 'The majority of teachers claimed that the main benefit to children's learning is the match which is facilitated between the experiences and activities provided and individual needs. This was especially emphasised in the case of pupils with special educational needs' (Hall *et al*., 1997, p 119). Teachers reported that their assessment (TA) gave them a better insight into the students' abilities and enabled teaching to be better focused. Teachers claimed that because of the assessment they were planning in greater depth and for short, medium and the longer term. It was also noted as significant that:

> In the course of the interviews teachers did not indicate that they were aware of or overly concerned with the accountability dimension or TA. Rather, the impression they created was one in which they themselves, and indeed the school in general, were still not accountable in any sense except to the children. (Hall *et al*., 1997, p 120)

Teachers were adapting their practices in line with the assessment requirements and the consequences were enhanced learning opportunities. However, there was a number of negative aspects: for instance, a focus on a single year (Year 2), rather than the whole key stage. Given the tradition of one teacher being responsible for one class of children for one school year, this tended to put the focus on the Year 2 teacher. This focus needs to be extended to other teachers in order to include the whole school 'and thereby create a greater degree of professional trust among teachers' (Hall *et al*., 1997, p 121).

Main points relating to the impact on students of the use of teachers' summative assessment for internal school purposes

There is high weight evidence of the following:

- Feedback from earlier assessment impacts on the effort that students apply in further tasks of the same kind; effort is motivated by non-judgemental feedback that gives information about how to improve (Brookhart and DeVoge, 1999; Carter, 1997/8).

- The way in which teachers present class assessment activities affects students' orientation to learning goals or performance goals (Brookhart and DeVoge, 1999).

- Changing teachers assessment practices to include processes and explanations leads to better student learning (Flexer *et al*., 1995).

- Using grades as rewards and punishments is harmful to students' learning by encouraging extrinsic motivation (Iredale, 1990; Pilcher, 1994).

There is medium weight evidence of the following:

- Teachers' own unguided grades are influenced by non-achievement factors such as students' behaviour, effort, attendance and disadvantaging some students. (Bennett *et al*., 1993; Cizek *et al*., 1995/6).
- The introduction of teachers' assessment related to levels of the National Curriculum in England and Wales was perceived by teachers as having a positive impact on students' learning experiences (Hall *et al*., 1997).

4.2.2 Evidence from studies reporting impact of teachers' summative assessment practice on teachers and teaching and the curriculum

Sixteen studies provide evidence of impact of teachers' summative assessment on teachers or teaching, including two that also gave some evidence of impact on the curriculum. As Table 4.3 indicates, there are only three studies providing high weight evidence in relation to the review question, 11 provided evidence of medium weight and two evidence of low weight.

Table 4.3: Studies providing information of the impact of teachers' summative assessment on teachers, teaching and the curriculum

Study	Age of students assessed (years)	Type of study	Use(s) of result	Overall weight of evidence
Yung (2002)	17–20	Description	External for certification (high stakes for student)	High
Flexer *et al.* (1995)	5–10	Evaluation: researcher-manipulated	Formative and summative Internal (for grading, in-school records, reporting to parents) Research	High
Johnston *et al.* (1993)	5–!0 11–16 17–20	Exploration of relationships	Internal (for grading, in-school records, reporting to parents)	High
Abbott *et al.* (1994)	5 – 10	Evaluation: Naturally occurring	External for accountability	Medium
Bennett *et al.* (1992)	5–10	Evaluation: naturally occurring	External for accountability	Medium
Cizek *et al.* (1995/6)	5–10 11–16 17–20	Exploration of relationships	Internal (for grading, in-school records, reporting to parents)	Medium
Gipps and Clarke (1998)	5–10 11–16	Evaluation: naturally occurring	Internal (for grading, in-school records, reporting to parents)	Medium
Hall and Harding (2002)	5–10	Evaluation: naturally occurring	Internal (for grading, in-school records, reporting to parents)	Medium
Hall *et al.* (1997)	5–10	Evaluation: naturally	Internal (for grading, in-school records, reporting to parents)	Medium

		occurring		
Hill (2002)	5–10	Evaluation: naturally occurring	Formative and summative Internal (for grading, in-school records, reporting to parents)	Medium
Koretz *et al.* (1994)	5–10 11–16	Evaluation: naturally occurring	External for accountability	Medium
McCallum *et al.* (1993)	5–10	Evaluation: naturally occurring	Internal (for grading, in-school records, reporting to parents)	Medium
Valencia and Au (1997)	5–10 11–16	Evaluation: naturally occurring	Formative and summative Internal (for grading, in-school records, reporting to parents)	Medium
Whetton *et al.* (1991)	5–10	Description	Internal (for grading, in-school records, reporting to parents)	Medium
Hiebert and Davinroy (1993)	5–10	Evaluation: researcher-manipulated	Formative and summative Internal (for grading, in-school records, reporting to parents)	Low
Morgan (1996)	11–16	Description	External for certification (high stakes for student)	Low

Studies where the assessment has an impact on teachers/teaching and is for external purposes of certification and/or accountability

Of the studies of assessment used for external purposes, only one, by Yung (2002) provides high weight evidence of impact on teaches and teaching. This study examined how teachers adapted their practice to accommodate the requirements of the replacement of an external practical examination by a school-based scheme, the Teacher Assessment Scheme (TAS). The different approaches of three teachers, taken as cases, were related by the author to the teachers' professional consciousness (commitment to educational goals) and professional confidence (in interpreting requirements in a way that is consistent with educational goals). The impact of the external regulations for conducting the TAS on one teacher was to keep strictly to the letter of what was required; this was interpreted as protecting his own interests rather than those of his students. Where there was ambiguity or scope for interpreting the regulations in different ways, this teacher chose to 'play safe', did not give students the benefit of any doubt and put himself in the role of a 'policeman'. By contrast, a teacher who put the importance of a rich learning opportunity for his student before his own interests, looked for the maximum flexibility in applying the regulations, allowing the assessment to interfere as little as possible with his usual relationship with students. Although this says as much about the impact of teachers' values on teacher-based assessment as about how the assessment affects teaching, the relationship is clearly two-way.

Of relevance to teachers' judgements in the context of external examinations is the study by Morgan (1996) of how teachers score coursework in mathematics for the GCSE in England. This suggests that, while coursework is intended to allow creative and unusual responses to problems, there can be pressure on teachers to avoid non-routine responses in the interests of reliable marking. The evidence, of low weight for this review, came from asking teachers to talk aloud as they read and assessed examples of coursework. Morgan concludes:

> …the examination process itself thus discourages creative or unusual ways of thinking. Exciting, innovative, creative mathematician-students may well work in non-routine ways and hence need to develop non-routine ways of communicating their work. While these characteristics may lead some teachers to value the work, they are also likely to give rise to conflicts both between different teachers and between the different positions potentially adopted by a single teacher. Such conflicts jeopardise the assessment system's requirements for reliability and simplicity. Teachers must, therefore, be under pressure to avoid them while preparing their students for the coursework examination (Morgan, 1996, p 372).

The study by Koretz *et al.* (1994) provides evidence of medium weight for this review, although for the review of reliability and validity of teachers' assessment (Harlen, 2004) the relevant evidence was rated as high weight. Koretz *et al.* (1994) studied the Vermont portfolio system, which had two purposes: 'to provide high quality data about students' achievement (in this case sufficient to permit comparisons of schools or districts) and to induce improvements in instruction' (Koretz *et al*., 1994, p 5). While one outcome of the study was a report on the reliability of scoring portfolios, the researchers also report on the impact on teachers and teaching. For the latter purpose, interviews were conducted with teachers and principals in a stratified random sample of about 80 schools. An anonymous questionnaire was also administered to all teachers participating in the mathematics portfolio programme. The findings were that teachers and principals characterised the programme as a 'worthwhile burden'. Participation in the programme demanded a lot of time and resources, and imposed considerable stress; teachers reported spending an average of 30 hours a month working on mathematics portfolios. Administrators had to commit large resources as well, such as using funds for substitute teachers. As well as time demands, there were difficulties finding appropriate tasks. At the same time:

> many educators found the program a powerful and positive influence on instruction. Mathematics teachers reported devoting more time to problem solving and communication. …Many noted changes in instructional practices as well; for example, about half reported an increase in the time students spent working in pairs or small groups. About half of the teachers reported that they had become more positive about mathematics but fewer (43% grade 4 and 27% grade 8) reported improvement in students' attitudes. Perhaps the more telling sign of the positive regard educators had for the program was that by end of first year of implementation, principals in roughly half of our sample schools reported that they had expanded use of portfolios beyond the grades and subjects participating in the state assessment program (Koretz *et al*., 1994, p 6).

The researchers conclude that, although the Vermont programme had shown promising effects on instruction and modest improvement in measurement quality in mathematics, the basic lesson to be drawn is the need for modest expectation, patience and ongoing evaluation with innovative large-scale performance assessments as a tool of educational reform.

Two of the studies providing medium weight evidence of impact on teachers and teaching concerned the implementation of the National Curriculum Assessment (NCA) in England and Wales. Abbott *et al.* (1994), as noted earlier, studied the administration to 7-year-olds by teachers of externally prescribed assessment tasks. The individual administration was enormously time-consuming for teachers;

this was the main impact reported. Teachers thought that the assessment tasks were pointless and a waste of time, which would have been better used in teaching. They thought that they could obtain better information from observation in normal teaching interactions. These observations were made at an early stage in implementation of the NCA and the tasks were subsequently substantially changed. Evidence from other studies suggests that, as time went on, teachers began to find useful information from the tests, which they could use in adapting their teaching, but at this early point the teachers were 'bogged down in organising materials, exploring, questioning and watching' and the main impact was in the use of time (Abbott *et al.,* 1994, p 167/8).

Many of these points were, not surprisingly, echoed in the study by Bennett *et al.* (1992), in which data were also collected in 1991 about the NCA for 7-year-olds in England. In this case, the study comprised a national questionnaire survey of primary teachers in which data were collected about teachers' subject knowledge and perceived competence to teach the National Curriculum as well as about NCA. For the TA the overall picture was:

> that a large proportion of teachers felt that time was substantially reduced for normal classwork, and that although more teachers felt that they were under additional stress (86%), class organization was largely unaltered (43%). A large proportion felt there was some value in this form of assessment for the planning of future lessons. There was some consciousness-raising and information learning regarding judging levels of attainment and acquiring knowledge about assessment techniques, but nearly half the teachers reported they did not feel professionally more competent as a function of their conducting TAs (Bennett *et al*., 1992, p 68).

For the standard tasks (SATs), most teachers felt time for normal classwork was even more reduced than for TA and most registered additional stress. As Abbott *et al*. (1994) report, over half the teachers found discipline a problem for at least some of the time during SAT administration. TA had little or no effect on the students.

Comparing TA with SATs, teachers preferred TA. However, only a minority thought they gained no new information about their students from the SATs. Bennett *et al.* (1992) pointed out that that there is conflicting evidence on this point, as the vast majority of teachers in another survey claimed that SATs were not telling them anything new (NUT, 1991). 'Notwithstanding, many primary heads, in their free-form responses, felt SATs were substantially a waste of time in their present form' (Bennett *et al.*, 1992, p 77). The authors infer from their study that 'the nature and extent of the support given in the last two years is inadequate to the task of successfully implementing major innovations on a national scale' (Bennett *et al.*, 1992, p 77).

Main points about the impact on teachers and teaching of summative assessment by teachers for external purposes

There is high weight evidence of the following:

- Teachers vary in how they respond to being given the role of assessor and the approach they take to interpreting external assessment criteria; strict adherence to the regulations leads them to be less concerned with students as individuals (Morgan, 1996; Yung, 2002).

There is medium weight evidence of the following:

- The impact on teaching of external assessment requirements depends on the value that teachers find in the information they gain about their students through the assessment (Abbott *et al*., 1994; Bennett *et al*., 1992; Koretz *et al*., 1994).
- Assessment for external purposes adversely affects teachers when it is seen as taking up too much time from teaching (Abbott *et al*., 1994; Bennett *et al*., 1992).

Studies where the assessment has an impact on teachers/teaching and is for internal purposes: grading, school records or reporting to parents

The high weight study by Flexer *et al.* (1995) report the effect of professional development on teachers' 'beliefs and practices about curriculum, instruction and assessment' (p 3). The change in teachers' practice in using more hands-on activities and asking students for explanations has already been mentioned. These changes followed from teachers being made aware through the new assessment techniques that they tried of how to make students think more deeply and become more engaged with a task. These ideas were applied in their teaching, thus providing more opportunities for systematic observation of students' learning. 'In short, the introduction of performance assessment provided teachers with richer instructional goals than the mere computation and raised their expectations of what their students can accomplish in mathematics and what they could learn about their students' (Flexer *et al*., 1995, p 33).

The researchers speculated about the changes that may have occurred in teachers' beliefs about learning and teaching mathematics:

> While we did not try to change beliefs directly, we know we affected beliefs through change in practice. There is no doubt that changes in beliefs alter practice, but it is also the case that shifts in practice may lead to shifts in belief, which can, in turn, further affect practice. In this study the changes that teachers made were likely to be changes in practice. We saw teachers, whose students gained greater understanding of multiplication from many hands-on activities, change their belief about how to teach multiplication. As teachers got positive feedback from students about changes they had made in instruction and assessment, they were encouraged to attempt further changes. In other words, changes in beliefs and changes in practices appear to be mutually reinforcing. While this cycle appeared to lead to, for some, a fundamental change in instructional and assessment *practic*e, it is not yet clear whether it also changed their *beliefs* about instruction and assessment. (Flexer *et al*., 1995, p 34)

The researchers claimed, however, that the changes that occurred resulted 'not from anything we told teachers to do, but from their experiences with the ways performance assessment improved their classrooms' (Flexer *et al*., 1995, p 35).

Johnston *et al.* (1993), in another study of high weight evidence, examined the relationship between teachers' assessment of students' learning of literacy and their own knowledge, values and teaching situation. The study involved teachers from districts where there were different degrees of control over the reading programme and the assessment of literacy by teachers. Of the 50 teachers who volunteered to be interviewed, 21 were from a district with 'low control', where teachers were encouraged not to use a basal reading scheme; 13 were from a district with 'high control', where there was strict enforcement of the basal reading scheme; and 16 teachers were from three districts of 'medium control' – somewhere between the two extremes. The sample included teachers of all Grades, from 1 to 12. In the high control district, teachers were required to use a basal reading programme which involved regular testing. 'Teachers were required to turn in end-of-unit tests to the district office to be scored. Their students could be 'spot tested' by administrators, with 24-hour notice, on skills they should have covered in the basal at a given point in the year' (Johnston *et al*., 1993, p 94).

The teacher interviews included a request to choose a student and describe his or her literacy development in as much detail as possible. The researchers found that descriptions by elementary teachers were more detailed than those of secondary school teachers, which was not surprising given that the upper grade teachers were keeping track of 100 or so students seen for only 45 minutes a day. However, in the high control district, elementary teachers also gave brief descriptions. Descriptions also differed in the extent to which they used 'subjective' language as opposed to 'objective' language. The more controlling the context, the more distancing terms such as 'skill' and 'mastery' were used. In low control context, teachers began with a personal introduction to the student as an individual, not with levels; the researchers found their descriptions more personal and more complex.

In relation to the teachers' assessment strategies, the methods for assessing children's literacy development reflected their goals and values but also the constraints under which they worked. In the low-control district, records were based on classroom observations, individual conferences with students, students' writing in journals, and tests, particularly informal, individualised ones. In the high control district, 'keeping track of students' development through observation of students' behaviour and comments was less common…serious testing was emphasised in every grade in the high-control context' (Johnston *et al*., 1993, p 103). The researchers commented, however, that although they used the tests as required, some of the teachers were sceptical about their usefulness.

The authors conclude the following:

> When tests were stressed by the district, teachers' descriptions of development emphasised tests, competitive attainment, and test-like language, and they turned to tests for feedback about students and themselves. In high-control contexts, teachers' assessment of students' development was relatively scant, reflecting a lack of detailed knowledge of the students and a less personally involving relationship (Johnston *et al*., 1993, p 113).

Thus teachers knew less about their students and focused their teaching on what was required for the tests. Johnston *et al.* (1993) hint that this could have an unintended impact on some students. They refer to the work of Broikou (1992) who found that teachers who routinely referred many students to special education gave brief and standardised descriptions of students' literacy development, whereas those who referred few students gave more extensive and individualised descriptions. 'In our study, technical and bureaucratic control of teachers' instructional practice, partly through the use of tests, produced unelaborated descriptions characteristic of teachers who refer many students'

(Johnston *et al*., 1993, p 114). In general, they conclude that students are not served well by conditions that constrain teachers' freedom to assess their students in productive ways.

Five of the studies giving medium weight evidence in this section concerned the teachers' assessment required by the National Curriculum Assessment (NCA) in England. This assessment, which is additional to the administration by teachers of externally set standard tests or tasks, at ages 7, 11 and 14, is reported externally but information used for school accountability, based almost entirely on the test scores. Thus there should be no high stakes attached to teachers' assessment and the main use of the results is internal to the school and for reporting to parents. For this reason these studies are included in this section. All, except Gipps and Clarke (1998), dealt with the assessment at the end of the second year of school (Year 2). Gipps and Clarke (1998) were concerned with assessments at age 7, 11 and 14. (Two other studies of NCA – by Abbott *et al.* (1994) and by Bennett *et al.* (1992) – involved both statutory teachers' assessment and national tests so have been discussed in the section on external use.)

The study by Hall *et al.* (1997), discussed earlier, gave evidence of a positive impact of the process of teachers' assessment on students' learning and their learning opportunities. They note that, as well as a particularly strong impact on teachers' planning, the assessment by teachers was seen as having an influence on all aspects of the implementation of the National Curriculum, from curriculum planning before the school year begins to summative, individualised reporting on each child at the end of the school year. In this sense, it seems that attempts were made to integrate assessment into the act of teaching and not merely add it on to satisfy official requirements. Schools were developing assessment policies and, where these existed, there was consistency between what the teachers reported as their practices and the policy. There was some concern, however, that the focus on assessment of the core subjects – English, mathematics and science – meant that teachers spent more time and energy teaching these subjects, resulting in fewer opportunities to develop children's learning in other subjects.

In a later study by Hall and Harding (2002), further attention was given to the impact of assessment on teachers' planning, particularly collaborative planning and the development of 'a community of practice'. This study was conducted after the change in the requirements for teachers' assessment within the NCA, which took place in 1993. In 1991 teachers had to assess their students against a number of criteria (statements of attainment) but, after a review in 1993, these statements were replaced by 'level descriptions' against which teachers judged a students' performance using a 'best fit' approach. The aim of Hall and Harding's study was to assess the extent to which a community of assessment practice is evident in schools in relation to the use of level descriptions. 'Assessment community' refers to a shared understanding among staff of the goals of National Curriculum assessment; a shared set of processes for the pursuit of these goals; and a common usage of a range of tools such as portfolios and exemplification materials to help staff with their assessment tasks.

All the Year 2 teachers and assessment co-ordinators in six schools took part in the study. The teachers and two local authority advisers in assessment were interviewed and classrooms were observed over a two-year period, 1998 and 1999. The main focus of the data analysis was to identify approaches to teachers' assessment adopted at the school level. They identified two main approaches, which they called ‘collaborative’ and ‘individualistic’. The former exhibited many of the characteristics of ‘an assessment community’, whereas teachers in the latter tended to work largely in isolation from their colleagues. They cited ‘abundant evidence’ of an association between the quality of professional relationships among teachers and the quality of teaching and learning in the classroom. Thus there is a potential for increasing quality of the students’ learning experiences through building professional cultures among primary teachers. But the extent of the positive impact was severely limited by the time allocated to it and the competing pressure of other centrally imposed initiatives The fact that funding was not made available for teachers to moderate their teachers’ assessment results served to tell teachers that the results of the external testing programme were prioritised over teachers’ assessment (TA). ‘The fact that TA, more than most other recent initiatives introduced into schools, depends on teachers exercising their professional judgement meant that teacher professionalism was enhanced and affirmed accordingly. Its diminished status, therefore, threatens that sense of professionalism’ (Hall and Harding, 2002. p12).

Whetton *et al.* (1991) report evidence of the methods used in teachers’ assessment in the first full year of NCA for 7-year-olds. This study was part of an ongoing evaluation of NCA. Most of the findings are details of types of support experienced by teachers. Impact on teaching was reported as arising when teachers were required to assess aspects of the curriculum that had not been covered. In the spring term, teachers struggled to teach and assess all aspects not previously covered. The result was increased workload, distortion of the curriculum and anxiety. In common with other studies of the early years of the NCA, many teachers in this study commented upon the excessive workload they had to undertake to complete their teachers’ assessment. Case studies showed that much of the workload was caused by inefficient systems for collecting and recording evidence of attainment. Unfamiliarity with NCA also contributed. Those schools which had been involved in the 1990 pilot approached the tasks with greater calm and confidence. Their assessment and recording policies tended to work more efficiently, resulting in less stress and a more manageable workload.

McCallum *et al.* (1993) looked at the models implicit in the way that individual teachers went about their teachers’ assessment. They collected data from interviews with the head teachers and Year 2 teachers in 32 schools and conducted intensive case studies in six schools. The data showed wide variation in many aspects of teachers’ assessment practice. The researchers were able to discern three main approaches. They noted that, since no approach to teachers’ assessment had been offered centrally, it was not surprising that teachers devised their own and that these should reflect their espoused views of teaching and learning. ‘What is particularly interesting to us as researchers is the relation between teaching, learning and assessment in the teachers’ practice; the link between assessment and learning is a crucial one but is not generally widely addressed’ (McCallum *et al.*, 1993, p 323).

Gipps and Clarke (1998) conducted a small scale evaluation of teachers’ perceptions of the assessment by teachers (TA) required by the NCA, the use being made of the support materials for TA provided by the national agency for assessment and the further help that teachers needed. The data were collected by questionnaire sent to the assessment co-ordinators and/or heads of English, mathematics and science departments in 300 primary and secondary schools. Interviews were conducted in 24 schools.

The findings of relevance to this review, contributing evidence of medium weight, concerned the effect on teachers of procedures adopted to conduct and standardise TA within the schools. For standardisation, most schools used meetings as the main methods for ensuring consistency of interpretation of level descriptions. Some secondary departments used a reference portfolio of work or followed a GCSE model. Standardisation meetings were held with greater frequency in primary than in secondary schools, but all found them manageable. For collecting information, the most common method was regular assessments and classroom tests. Primary school co-ordinators and heads of English departments reported less use of tests than other teachers and were more likely to involve students in self-assessment. In general, primary teachers spent much longer assigning levels than secondary teachers. The value that teachers found in the process was reflected in their comments and the high level of support for TA to be reported: that is more valid than giving tests; maintains a broad curriculum; encourages teachers to keep track of students throughout the year; is the core of learning; raises students' awareness and their profile of achievement; and enables all students, especially those not good at taking tests, to be fairly assessed. In summary, Gipps and Clarke (1998) report:

> Assessment was recognised as making a crucial contribution to teaching and learning. There were a few references to assessment taking up teaching time; however, both formative and summative teacher assessment were recognised as being essential. There were some negative references, particularly amongst secondary teachers, to the over-emphasis given to the publication of results and, in particular, league tables.
>
> End-of-key stage teacher assessment was recognised as giving a more rounded view of a pupil's achievements and capabilities. It was also regarded as leading to the development of the skills needed for ongoing formative assessment and, certainly by primary teachers, as extending the teacher's knowledge of the curriculum.
>
> Despite the workload involved, all of the teachers valued teacher assessment and believed it should remain a part of end-of-key stage National Curriculum assessment (Gipps and Clarke, 1998, p 7).

Valencia and Au (1997) report a study of unusual design for investigating the use of portfolios for assessing literacy. Teachers at two sites, where different portfolio assessments were in operation, were interviewed before a series of cross-site meetings. In these two-to-three day meetings, they observed in each other's classrooms and then, in groups, reviewed and evaluated portfolios from each site, using first the criteria developed and used at one and then the other site. Finally, they developed common criteria and used these to evaluate the portfolios. After these meetings, they were again interviewed. The authors report the teachers involved as being 'struck as much by the similarities in their philosophies and instructional emphasis as by the differences in the portfolio contents' (Valencia and Au, 1997, p 18). Their content analyses of the portfolios found that they contained high quality authentic samples and records of students' reading and writing. Reading outcomes and discussion were more difficult to document. There was a marked consistency in the teachers' marking of each others' portfolios, which the researchers attributed to shared conceptual understanding, further enhanced through classroom visits. Equally important, they pointed out, was the low stakes nature of the project.

The impact on teachers is illustrated by this example:

> At the conclusion of the project, Sue noted that she had started being more specific with her students about the kinds of evidence needed to document their involvement with the writing process.... She had her students evaluate their own writing using an evaluation rubric as a means of teaching them about standards for literacy performance. She had also become more aware in her own classroom of the quality of students' responses during literature discussions. A portfolio content analysis served to verify that Sue had succeeded in making several changes. (Valencia and Au, 1997, p 26)

The authors conclude that, as the teachers closely examined students' work from their own and other classrooms, they clarified the meaning of learning outcomes and learned to interpret student performance based on multiple forms of evidence.

> Working collaboratively and discussing portfolio artifacts encourage teachers to re-examine their knowledge, assumptions and misconceptions about teaching and assessment practices. Conversely, teachers' understandings of curriculum and instruction are reflected in portfolio evidence. We can see their strengths and weaknesses, their priorities, and the opportunities they provide for their students. As teachers grow and change through professional development, their portfolio evidence begins to change as well. This slow, iterative process is likely to produce meaningful sustainable changes in both portfolio evidence and in underlying classroom instruction. (Valencia and Au, 1997, p 30)

Cizek *et al.* (1995/6) report a contrasting situation in the practice of teachers in grading students when this is unguided by criteria, grading policies or evidence collected, as in a portfolio. The study, outlined earlier, reports wide differences in the information used and the process of arriving at grades. The main common feature was that 'with discernible regularity, teachers appeared to structure their assessment practices and combine formal and informal assessment information in ways that were most likely to result in a higher grade for their students' (Cizek *et al.*, 1995/6, p 175). There was little evidence of the grades or the records of them being used to influence teaching and learning. The researcher recommended that districts begin to consider, establish and disseminate information that would provide guidance to teachers about desirable assessments and grading practices. It was also recommended that schools should more actively pursue engendering cultures of collaborative practice, especially related to assessment. Without this consistency, 'it is not clear that any interested group – administrators, teachers, parents or even students and teachers themselves – can confidently glean the meaning of the grades students receive' (Cizek *et al.*, 1995/6, p 175).

Hill (2002) reports how some New Zealand teachers responded to new assessment requirements introduced in 1993. Although the officially given purpose was to improve learning and teaching, the author argues that it 'blurred the boundary between the primary purpose of emphasising learning and the purpose of accountability.' (Hill, 2002, p 116). The study involved interviews of 12 primary teachers in two case study schools over two years (1996 and 1997) as they implemented new assessment policies. As in the study of McCallum *et al.* (1993), Hill found varying strategies used by teachers, but all tended to separate assessment from teaching. Teachers frequently used special tasks to check off achievement against objectives. Formative assessment was framed as teaching, not assessment. Hill described their approach as 'driven from above'. She concluded that the potential for school-based assessment to inform teaching and learning was restricted by the use of checklists and 'quasi grading systems that encouraged teachers to dot, slash, cross against achievement objectives' (Hill, 2002, p 121). What is needed is for teachers to internalise the nature of progressions and work with these in mind.

This conclusion was also supported in the study by Hiebert and Davinroy (1993), concerning the introduction of new forms of assessment to teachers in three schools. This study adds evidence of low weight that, once teachers themselves understand the progression in the skills to be developed by students, they begin to expect more of their students.

Main points about impact on teachers, teaching and the curriculum of summative assessment for internal purposes

There is high weight evidence of the following:

- The introduction of assessment techniques that require students to think more deeply leads to changes in teaching that extend the range of students' learning experiences (Flexer *et al.*, 1995).

- Close external control of teacher assessment inhibits teachers gaining detailed knowledge of their students (Johnston *et al.*, 1993).

There is medium weight evidence of the following:

- When teachers' assessment is built into teachers' planning, the process has a positive impact on teaching and learning. This impact is further enhanced by professional collaboration at the school level (Hall *et al.*, 1997; Hall and Harding, 2002).

- Assessment by teachers indicates where learning opportunities for their students need to be extended (Whetton *et al.*, 1991; Valencia and Au, 1997).

- In a low stakes context, the process of summative assessment by teachers helps them to clarify the meaning of learning outcomes (Valencia and Au, 1997).

- The value of teachers' summative assessment for potential users depends on teachers internalising the nature of progression in relation to the learning goals (Cizek *et al.*, 1995/6; Hill, 2002).

4.3 Synthesis of evidence: subsidiary review question

What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes?

Summative assessment is essentially a form of communication. While the message communicated – the grades or mark – is likely to affect students, particularly when important decisions depend on it, the focus in this review is on the process of the assessment. Any summative assessment involves decisions about the domain of knowledge, skills and other attributes to be assessed; gathering a valid sample of student behaviour in the domain; criteria for judging the sample of behaviour; procedures for applying the criteria; and procedures for reporting and communicating to those who need to be informed about the achievement of the students (based on Harlen, 2004, p 83). Looking across the studies in the in-depth review, there are six main conditions which appear to affect the process of assessment when it is carried out by teachers and therefore the kind of impact that this process has or is likely to have. In some cases, these apply to impact on students directly and in others may affect students indirectly through impact on teachers and teaching.

The conditions identified are as follows:

- the balance between the demands on teachers and the benefits they perceive from conducting summative assessment of their students
- the extent to which high stakes are attached to the assessment
- the existence and use of assessment criteria or policies to guide teachers' judgements
- the use of grades as rewards
- whether the assessment is based on 'summing up' or 'checking up'
- Professional development opportunities for teachers to share and develop their understanding of assessment procedures and how they relate to teaching and learning

4.3.1 The balance between demands and benefits

When teachers are required to administer externally devised tasks or to use externally imposed criteria in assessing students' ongoing work, there is inevitably impact on teachers' practice, if only in determining how some of their time is used. The extent to which the impact on practice is positive depends on the balance between the value that teachers may find in the required processes of assessment and the constraints imposed. When the National Curriculum Assessment of 7-year-olds in England and Wales was first introduced in 1990, teachers were required to administer standard externally devised tasks as well as to conduct their own assessment of students' achievement. In the first year of the full implementation, a large proportion of teachers reported that these tasks occupied a great deal of time, disrupted their classroom and were thought to be 'a waste of time' because they learned nothing new about their students through the process (Bennett *et al.*, 1992; Abbott *et al.*, 1994). The demands on teachers' time of conducting their own assessments were also large, but conducting these assessments was not as disruptive of the classroom and brought benefits in terms of the information provided about the students (Gipps and Clarke, 1998). Whetton *et al.* (1991) also report that, with experience, teachers adopted more efficient approaches to the teachers' assessment, bringing less stress and making the process less time-consuming. However, there was no opportunity for teachers to change and assimilate into their practice the standardised tasks. (These tasks were changed by the National Assessment Agency in response to the problems that had been caused to teachers.)

Koretz *et al.* (1994) found that the portfolio assessment introduced in Vermont was felt to be very time-consuming by teachers, and associated with considerable stress; nevertheless, the value to schools through the positive impact on teaching and on students' attitudes was considered to make it 'a worthwhile burden'.

The study by Flexer *et al.* (1995) gave high weight evidence that providing teachers with assessment tasks and with professional development in assessment related to reforms in mathematics education had a positive influence on teachers' understanding of learning in mathematics and how they needed to change their teaching practice to reach new goals. In this intervention, which was spread over two years, teachers developed their understanding through trying new ways of assessing and seeing the response of students for themselves. Flexer *et al.* (1995) note that teachers were able to take ownership of the procedures and to see the assessment as essential to teaching. The implications from this study are that teachers need the opportunity to try new ideas and to judge for themselves the value; also, it takes time for this to happen.

When new assessment arrangements require a considerable change from previous practice, as in the case of the NCA, there is need for support and time to accommodate to the demands. Gipps and Clarke (1998) found evidence that, by the mid-1990s, teachers were finding that the teachers' assessment required by the NCA was not only helpful but essential and 'the core of the learning process'.

4.3.2 The high stakes attached to the outcome of the assessment process

There is evidence that extent and type of impact on teachers' practice from external assessment requirements is related to the stakes attached to the outcome. When the stakes are low, there are greater opportunities for positive impact than when they are high. Thus Valencia and Au (1997) note that the low stakes nature of their work with teachers was 'integral' to the professional collegiality that developed during its course. The positive impacts on teachers and their teaching of the professional development reported by Flexer *et al.* (1995) was in the context of the normal standardised testing programme being waived, for the schools involved, by the districts where the study took place.

Where there was high stakes use of the assessment results, there was evidence of a reduced positive impact of the teachers' assessment on both students and teachers. Bullock *et al.* (2002) note that, in the context of teacher-assessed coursework, there was greater flexibility allowed in the procedures than in the teachers' practices. It was the pressure on teachers to enable students to reach high grades that appeared to encourage teachers to keep to more routine tasks, for which they could coach students in the specific skills required. In turn, this restricted the opportunities of students to take full advantage of the opportunities the coursework offered for developing higher-level thinking and prevented students from taking ownership of the coursework.

Not all teachers respond to high stakes in the same way. Yung's (2002) study found that the extent to which teachers held strictly to a narrow interpretation of the assessment regulation in a practical biology examination was a function of their desire to protect either their own, or their students interests. The teachers who wanted to be seen as applying the regulations precisely, and those who interpreted them more flexibly in order to give students higher grades, put their own interests first. Other teachers who used the regulations flexibly did this to ensure a worthwhile experience for their students. Morgan (1996) also found a difference among teachers according to whether they rigidly applied criteria to a student's work, or whether they tried to understand what the student was attempting to convey.

Hall and Harding (2002) report a change in primary teachers' assessment in the NCA in England during the 1990s on account of the pressure to raise standards. They quote a teacher as saying that 'Up to now our TA has been "If you're not sure (about a higher level), then don't give it" and now you're having to think, "Well, if you're not sure, we'd better give it"' (Hall and Harding, 2002, p 12). Hall and Harding also point out that this shift in practice became more likely as less time was given to school-based moderation, discussion of assessment and to the development of 'communities of assessment practice' due to priorities having shifted to new initiatives. Thus teachers were making decisions as individuals rather than in a collaborative context.

4.3.3 The existence and use of external criteria or policies to guide teachers' assessment

Across the studies there is sharp contrast between summative assessment processes that are guided by policies and shared criteria, and those where individual teachers decide for themselves the basis of their judgements. The studies by Bennett *et al.* (1993), Pilcher (1994) and Cizek *et al.* (1995/6), all of grading practices in the USA, show that teachers, unguided by generally agreed policies on assessment, gave summative grades that are based on some combination of achievement in the subject and non-academic behaviour. This practice contrasts with practice of teachers in England since the introduction of the NCA. McCallum *et al.* (1993) and Hall *et al.* (1997) show that teachers gather evidence in different ways, but they all apply the same criteria in arriving at their judgement of the level at which their students are working. The difference in these practices clearly affects the product of the assessment – the grade or level – and the meaning it has for other teachers, parents and students. The high weight evidence provided by Pilcher's (1994) study, for instance, found that parents did not interpret grades as teachers assigned them, but perceived grades to reflect achievement levels. This is not to say that parents in England interpret levels in the same way as teachers, unless they have access to the National Curriculum levels, but at least levels from different teachers have much the same meaning.

Beyond the uncertain meaning of the outcome, however, the process of arriving at the outcome reflects, and impacts upon, what is given attention in teaching. Teachers value students making an effort, showing good behaviour and participating well in class. By incorporating all of these into a single grade, they are recognising the value of many different educational outcomes, but, in doing so, are not communicating anything of meaning about any of them. Combining a variety of different kinds of educational outcomes into one grade means that none is communicated effectively. One impact on students and their learning experience may well be for them to be regarded as less able than their achievement suggests, as in the case of young boys (Bennett *et al.*, 1993). Other impacts result from using grades as rewards and punishments, as discussed in the following sections.

4.3.4 The use of grades as rewards and punishments

The study by Brookhart and DeVoge (1999) provides empirical evidence that students' self-efficacy in relation to particular tasks is influenced by their achievement in tasks of similar kinds undertaken previously. Students obtain feedback from the assessment of their work which has a key role in determining the effort they will apply to new tasks. Brookhart and DeVoge (1999) notes that the form in which the feedback is given, and not just whether it is favourable or not, is important. Judgemental feedback, in the form of a mark or grade, may have a positive or negative impact, depending on whether the grade is high or low (although sometimes a high grade leads to less effort since the student is confident that the task is easy). However, non-judgemental feedback, that gives information about how the work can be improved, is likely to encourage effort on future occasions in all cases.

The evidence in some studies reviewed here is that, far from trying to reduce the impact of grades on students' motivation, teachers may use them as rewards or punishments. In these circumstances, students are encouraged to be motivated to apply effort by the prospect of good grades rather than by the learning that is intended. Findings from the first review by the ALRSG (Harlen and Deakin Crick, 2002) suggest that this can lead to shallow learning and to students embracing performance goals rather than learning goals. This use of assessment can of course occur with external assessment as well as teachers' assessment, but the frequency with which teachers give grades increases the impact. Moreover, as Brookhart and Devoge (1999) point out, the way in which teachers present assessment tasks also impacts on motivation and goal orientation.

When teachers' grades combine non-academic behaviour and effort with achievement as reported by Cizek *et al.* (1995/6) and Pilcher (1994), teachers are able to use grades as rewards or punishment for good or poor behaviour. As Bennett *et al.* (1993) and Pilcher (1994) report, this can mean students being given grades that do not reflect their academic achievement, with consequences for the expectations of them that may have a negative impact.

In summative assessment, the use of grades or levels at some point is inevitable, in the interests of efficient communication. It is essential, then, for all those involved to know the meaning of the grade or level. This is particularly important for students in order to avoid a negative impact on their motivation for learning. Bullock *et al.* (2002) and Iredale (1990) provide high weight evidence that students need more explicit descriptions and examples to enable them to understand the grades they are given.

4.3.5 Whether the assessment is based on 'summing up' or 'checking up'

The potential for the process of summative assessment by teachers to have an impact on teaching is clearly likely to be greater the more frequently teachers gather information. McCallum *et al.* (1993) found that, in the absence of guidance about how to conduct TA as part of the NCA in England and Wales, teachers adopted a range of strategies. Some recorded events on the spot, others relied on reflecting back on events from memory and others planned some tasks specifically for assessment purpose. The reports of Whetton *et al.* (1991) and Bennett *et al.* (1992), by reference to teachers giving extra attention to students in conducting TA, implied that special activities were set up and students were aware of being assessed. Hall *et al.* (1997) describe procedures used by teachers in which they gave planned specific assessment tasks but used the information in an iterative process in setting further tasks. Assuming that the assessment tasks were also learning tasks, the process of assessment was providing some information that was used formatively in adapting teaching. Only at a later stage did teachers reflect on the information gained in order to assign a level to each student. These procedures were spread over the school year and resulted, as Hall *et al.* (1997) report, in enhanced learning opportunities for the students. However, since the TA was only reported at the end of Year 2, the practice was not followed in Year 1.

The collection of work in portfolios, as in the Vermont programme studied by Koretz *et al.* (1994), involved attention to students' work over an extended time. By requiring that evidence of certain kinds of performance be included in the portfolios, the process had 'a powerful and positive influence on instruction' (Koretz *et al.*, 1994, p 6). By contrast, Johnston *et al.* (1993) report a negative impact on teachers' knowledge of students when teachers were required to administer regular tests, rather than use more informal methods of gathering information. No impact on teaching is reported in the studies of assessment conducted only at particular times (Bullock *et al.*, 2002; Iredale, 1990; Yung, 2002). However, Iredale (1990) and Bullock *et al.* (2002) report a positive impact on students due to the teachers' assessment being associated with less anxiety that formal examinations. Carter (1997/8) report a positive impact on students when they were given a role in the summative assessment of their own work.

4.3.6 Professional development opportunities for teachers to share and develop their understanding of assessment procedures and how they relate to teaching and learning

It is to be expected that professional development in assessment techniques improves teachers' assessment practice, but there is evidence of a wider impact on teachers. Flexer *et al.* (1995) found that this extended from teachers' beliefs about how students learn and about what it is important for students to be taught, to their teaching practice. These changes in teaching created more opportunities for students' learning and teachers reported that, at the end of a year of professional development, students were solving problems and giving explanations at a level that surprised many teachers. In relation to the time required, it is interesting to note that Hiebert and Davinroy (1993) found that a three-month period of staff development was not enough to determine whether their staff development sessions on assessment could change teachers' views of literacy and literacy instruction.

Gipps and Clarke (1998) found that teachers valued professional development, particularly in-school meetings, very highly; teachers found these sessions effective in facilitating understanding of the level descriptions in the National Curriculum. Hall *et al.* (1997) report, from interviews with Year 2 teachers in 1993, a sense of 'professional mistrust' of the TA results passed on by Year 1 teachers.

At that time, the training for TA was focused almost entirely on Year 2 teachers. At a later stage in the implementation of NCA, Hall and Harding (2002) note a distinction between the assessment practices and impacts of teachers' assessment in schools according to whether teachers worked collaboratively or individually on their assessment. In schools where teachers perceived themselves as working alone, the process of assessment had not led to more discussions among teachers about students' learning, as it had in collaborative schools.

The unusual cross-site professional development, which Valencia and Au (1997), initiated as part of their study of portfolio assessment, was associated with strong and beneficial collaboration among teachers. Looking closely at students' work from their own and others' classrooms clarified the meaning of important learning outcomes. Working collaboratively encouraged teachers to reflect on their assumptions and misconceptions about teaching and assessment practices.

Main points relating to the conditions and contexts that affect the nature and extent of the impact of using teachers' assessment for summative purposes

Each of these points is supported by some evidence of high weight (indicated by the **bold** references) as well as by evidence of medium and low weight.

- New assessment practices are likely to have a positive impact on teaching if teachers find them of value in helping them to learn more about their students and to develop their understanding of curriculum goals. Time to experience and develop some ownership of practices enhances their positive impact (**Flexer *et al.*, 1995**; Abbott *et al.*, 1994; Bennett *et al.,* 1992; Gipps and Clarke, 1998; Koretz *et al.*, 1994).

- When high stakes judgements are associated with teachers' assessment, one effect is for teachers to reduce assessment tasks to routine events and restrict students' opportunities for learning from them. High stakes encourages some teachers to give high grades where there is doubt, which may not be in the students' interests (**Bullock *et al.*, 2002; Yung, 2002**; Hall and Harding, 2002; Morgan, 1996).

- Shared criteria for assessing specific aspects of achievement lead to positive impact on students and on teaching; in the absence of such guidance, there is little positive impact on teaching and a potential negative impact on students (**Pilcher, 1994;** Bennett *et al.*, 1993; Cizek *et al.*, 1995/6; Hall *et al.*, 1997; McCallum *et al.*, 1993).

- The process that teachers use in setting assessment tasks and in grading, impacts on students' motivation for learning, particularly their goal orientation, when grades are used as rewards or punishments. The negative impact can be alleviated by ensuring that students have a firm understanding of assessment processes and criteria (**Brookhart and DeVoge, 1999; Bullock *et al.,* 2002;** Iredale, 1990; Stables, 1992).

- Summative assessment by teachers has a more positive impact on teachers and teaching when integrated into practice than when concentrated at certain occasions (**Bullock *et al.*, 2002; Johnston *et al.*, 1993;** Bennett *et al.,* 1993; Carter, 1997/8; Hall *et al.*, 1997; Iredale, 1990; Koretz *et al.,* 1994; McCallum *et al.*, 1993; Whetton *et al.*, 1991).

- Opportunities that enable teachers to share and develop their understanding of assessment procedures enables them to review their teaching practice and their view of students' learning and of subject goals. Such opportunities need to be sustained over time and preferably should include provision for teachers to work collaboratively across as well as within schools (**Flexer *et al.*, 1995**; Gipps and Clarke, 1998; Hall *et al*., 1997; Hall and Harding, 2002; Hiebert and Davinroy, 1993; Valencia and Au, 1997).

4.4. In-depth review: quality-assurance results

Data extraction for all 23 studies was carried out independently by at least two people, as described in section 2.3.5. Most differences were in the detail provided rather than in the main judgements. In general in reconciling data extractions, a more complete description was produced by combining aspects of detail provided by each extraction. In only two cases were there differences in study type, which had consequences for subsequent questions in EPPI-Reviewer. These concerned studies where qualitative methods were used in evaluating interventions and could be construed as exploration of relationships, in one case, or as description, in another. Differences in judgements of overall weight of evidence occurred in three cases. These were due to bias in favour of studies which, although of high quality in terms of methodology, did not give a great deal of information of relevance to the review question. These examples strengthen the case for the quality-assurance measures in the EPPI-Centre reviewing procedures.

4.5 Involvement of users in the review

The participation of users in conducting the review has been indicated in some detail in section 2.1. Those involved in keywording and data extraction included a teacher, professional developers and researchers, as well as research assistants. Despite the pressure on their time caused by the extra work, members of the Review Group reported unanimously that the 'hands-on' involvement was most valuable in understanding the processes of the review, the nature of evidence available from research and the issues surrounding the review question.

The opportunity for users to have a wider involvement in the discussion of implications of the review for policy, practice and research was curtailed by the reduced timescale for completing the review and the fact that the final stages were reached when most people were taking their summer holiday. However, an attempt was made to consult members of the Review Group by email on the findings and the implications of the review. This resulted in detailed comments and suggestions from six of the members and proved to be a very useful process for obtaining formative feedback.

5. FINDINGS AND IMPLICATIONS

This chapter begins by summarising the outcomes of the review that have been presented in Chapters 3 and 4. Some strengths and weaknesses of the current review are discussed and some implications of the review findings for policy, practice and research are identified.

5.1 Summary of principal findings

5.1.1 Identification of studies

The search for studies was carried out through a process of handsearching journals online and in the Graduate School of Education library, searching relevant electronic databases, and using citations and personal contacts. The total number of papers found was 343, of which 301 were excluded in either a one-stage or a two-stage screening, using inclusion and exclusion criteria. Full texts were obtained for 26 of the remaining 42 papers, from which a further two were excluded during keywording and one included paper was linked to another, as they were based on the same set of data. This left 23 studies included in both the systematic map and the in-depth review.

It was noted that, while 77% of the original 343 papers were found through searching online databases or through citations (i.e. a two-stage screening process), only 26% of the 23 included studies were from these sources. The one-stage searching provided 74% of the included studies.

5.1.2 Mapping of all included studies

The 23 studies included in the in-depth review were mapped in terms of the EPPI-Centre and review-specific keywords. All were written in the English language and 12 were conducted in England, nine in the United States and one each in New Zealand and Hong Kong.

All studies were concerned with students between the ages of 4 and 18. Eleven involved primary school students (aged 10 or below) only, six involved secondary students (aged 11 or above) only and five were concerned with both primary and secondary students. A slightly larger proportion of studies conducted in primary schools reported impact on teachers compared with those conducted in secondary schools. About 70% of studies in secondary schools and about 80% in primary schools were concerned with assessment of English; while 43% and 60% respectively were concerned with assessment of mathematics.

Twenty studies were classified as involving assessment of work as part of, or embedded in, regular activities. Three were classified as portfolios, two as projects and eight were either set externally or set by the teacher to external criteria. The most common use of the assessment in the studies was for internal school purposes, with four studies related to assessment for certification and another three to external purposes that had high stakes for the school.

5.1.3 Nature of studies selected for in-depth review

All the studies included in the map were also included in the in-depth review. Seven studies provide evidence of high weight for the review. Six of these provided information about impact on students; three also provide information about impact on teachers. Of the 12 studies providing evidence of medium weight, all except one provide evidence of impact on teachers, whilst five provided information of impact on students. Thus there is rather more evidence of high weight for impact on students.

5.1.4 Synthesis of findings from studies in in-depth review

Findings from the studies on impact on students have been summarised separately for external and internal purposes in section 4.2.1, and findings relating to impacts on teachers, teaching and the curriculum can be found in sections 4.2.1 and 4.2.2. These are brought together here so that evidence for all impacts can be considered together.

There is high weight evidence of the following:

- Older students respond positively to summative assessment by teachers of their coursework, finding the work motivation and being able to learn during the assessment process (Bullock *et al.*, 2002).

- Students need more help, in the form of better descriptions and examples, to understand the assessment criteria and what is expected of them in meeting these criteria (Bullock *et al.*, 2002; Iredale, 1990; Stables, 1992).

- The impact of summative teacher assessment on students depends on the high stakes use of the results (Yung, 2002).

- The impact of summative teachers assessment on students will be affected by the way teachers interpret their roles as assessors and by their orientation towards improving the quality of students' learning or maximising their marks (Bullock *et al.*, 2002; Yung, 2002).

- Feedback from earlier assessment impacts on the effort that students apply in further tasks of the same kind; effort is motivated by non-judgemental feedback that gives information about how to improve (Brookhart and DeVoge, 1999; Carter, 1997/8).

- The way in which teachers present classroom assessment activities may affect students' orientation to learning goals or performance goals (Brookhart and DeVoge, 1999).

- Changing teachers' assessment practices to include processes and explanations can lead to better student learning (Flexer *et al.*, 1995).

- Using grades as rewards and punishments is harmful to students' learning by promoting extrinsic motivation (Iredale, 1990; Pilcher, 1994).

- Teachers vary in how they respond to being given the role of assessor and the approach they take to interpreting external assessment criteria; strict adherence to the regulations leads them to be less concerned with students as individuals (Morgan, 1996; Yung, 2002).

- The introduction of assessment techniques that require students to think more deeply leads to changes in teaching that extend the range of students' learning experiences (Flexer *et al.*, 1995).

- Close external control of teacher assessment inhibits teachers gaining detailed knowledge of their students (Johnston *et al.*, 1993).

There is medium weight evidence of the following:

- Teachers consider that young students may not do their best work when constrained by an external task (Abbott *et al.*, 1994).

- Teachers' own unguided grades are influenced by non-achievement factors, such as students' behaviour, effort, attendance and disadvantaging some students (Bennett *et al.*, 1993; Cizek *et al.*, 1995/6).

- The introduction of teachers' assessment related to levels of the National Curriculum in England and Wales was perceived by teachers as having a positive impact on students' learning (Hall *et al.*, 1997).

- The impact on teaching of external assessment requirements depends on the value that teachers find in the information they gain about their students through the assessment (Abbott *et al.*, 1994; Bennett *et al.*, 1992; Koretz *et al.*, 1994).

- Assessment for external purposes adversely affects teachers when it is seen as taking up too much time from teaching (Abbott *et al.*, 1994; Bennett *et al.*, 1992).

- When teachers' assessment is built into teachers' planning, the process has a positive impact on teaching and learning. This impact is further enhanced by professional collaboration at the school level. (Hall *et al.*, 1997; Hall and Harding, 2002)

- Assessment by teachers indicates where learning opportunities for their students need to be extended (Valencia and Au, 1997; Whetton *et al.*, 1991).

- In a low stakes context, the process of summative assessment by teachers helps them to clarify the meaning of learning outcomes (Valencia and Au, 1997).

- The value of teachers' summative assessment of potential users depends on teachers internalising the nature of progression in relation to the learning goals (Cizek *et al.*, 1995/6; Hill, 2002).

Findings relating to the subsidiary review question (What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes?) were listed in Chapter 4. As indicated there, each is supported by evidence of high weight as well as evidence of medium weight.

- New assessment practices are likely to have a positive impact on teaching, if teachers find them of value in helping them to learn more about their students and to develop their understanding of curriculum goals. Time to experience and develop some ownership of practices enhances their positive impact (Abbott *et al.*, 1994; Bennett *et al.,*1992; Flexer *et al.*, 1995; Gipps and Clarke, 1998; Koretz *et al.*, 1994).

- When high stakes judgements are associated with teachers' assessment, one effect is for teachers to reduce assessment tasks to routine events and restrict students' opportunities for learning from them. High stakes encourages some teachers to give high grades where there is doubt, which may not be in the students' interests (Bullock *et al.*, 2002; Hall and Harding, 2002; Morgan, 1996; Yung, 2002).

- Shared criteria for assessing specific aspects of achievement lead to positive impact on students and on teaching; in the absence of such guidance there is little positive impact on teaching and a potential negative impact on students (Pilcher, 1994; McCallum *et al.*, 1993; Bennett *et al.*, 1993; Hall *et al.*, 1997; Cizek *et al.*, 1995/6).

- The process that teachers use in setting assessment tasks and in grading impacts on students' motivation for learning, particularly their goal orientation, when grades are used as rewards or punishments. The negative impact can be alleviated by ensuring that students have a firm understanding of assessment processes and criteria (Brookhart and DeVoge, 1999; Bullock *et al.*, 2002; Iredale, 1990; Stables, 1992).

- Summative assessment by teachers has a more positive impact on teachers and teaching when integrated into practice than when concentrated at a certain occasions (Bennett *et al.,*1993; Bullock *et al.*, 2002; Carter, 1997/8; Hall *et al.*, 1997; Iredale, 1990; Johnston *et al.*, 1993; Koretz *et al.,*1994; McCallum *et al.*, 1993; Whetton *et al.*, 1991).

Opportunities that enable teachers to share and develop their understanding of assessment procedures enables them to review their teaching practice and their view of students' learning and of subject goals. Such opportunities need to be sustained over time and preferably should include provision for teachers to work collaboratively across as well as within schools. (Flexer *et al.*, 1995; Gipps and Clarke, 1998; Hall *et al.*, 1997; Hall and Harding, 2002; Hiebert and Davinroy, 1993; Valencia and Au, 1997)

5.2 Strengths and limitations of this systematic review

5.2.1 Strengths

Most of the strengths of this review follow from its systematic and collaborative procedures. The identification at the start of precise research questions enables both (i) the search for studies and (ii) the selection and synthesis of evidence from the most relevant and dependable studies. Although no search can be comprehensive, the careful documentation of sources searched makes explicit the limits of the evidence base and enables later work to extend the range without duplication. This is further assisted by the identification of inclusion and exclusion criteria. These criteria enabled a range of study types and designs to be included. Although the main question in this review is one concerning impact, the evidence is not restricted to controlled trials. In the context of using teachers' assessment to summative purposes, a controlled trial, where the use of TA is compared with some other form of summative assessment and is the only independent variable, is hardly feasible.

However, there is no doubt that the quality of a good deal of the research found was quite low. Many studies report small-scale, short-term investigations where little attention had been paid to establishing the reliability and validity of either the data collection or data analysis. Therefore the procedures that enable decisions to be made about the weight that can be given to the evidence from particular studies is an essential safeguard to regarding all evidence as equally valuable, as can happen in a narrative review of research.

The collaborative nature of the procedures used is a further strength. No critical decisions have been taken by one person alone. At all stages, the decisions about the conduct of the review have been discussed among the Review Group members. At the more detailed level of inclusion and description of studies and data extraction, all decisions have been taken as a result of at least two people working first independently and then reaching a consensus about any differences. The quality assurance role of the EPPI-Centre staff has ensured that these processes were conducted reliably.

Finally this review has been carried out in collaboration with researchers involved in the second review of the Citizenship Education Research Strategy Group. The sharing not only of training, research assistance and technical skills, but also of ideas and approaches to common issues, has been a source of strength for the current review.

5.2.2 Limitations

In terms of search procedures, the search for studies by computer has been limited by the extent of journal articles available online. Although this is increasing, the access to back numbers is still limited in many cases. Online search had therefore to be supplemented by handsearch in the library. This means that access to studies from countries outside the UK was limited by the library holdings. Changes to the US ERIC database have made this a difficult source to use; it includes a large number of conference papers and local reports, which cannot be obtained in full text. Further, only studies written in English have been used in this review.

Just over 50% of studies selected for in-depth data extraction were from the UK. However, as the rationale for this study is related to practice and policy decisions in the UK, this figure is perhaps not surprising. UK researchers have been drawn to this area of work and funders have provided support in view of the particular concerns surrounding summative assessment reflecting the growing interest (mentioned in the introduction to the Background section) in giving teachers' assessment a greater role in summative assessment.

The restriction to assessment of school students means that potentially relevant evidence from higher and further education, where teachers' assessment is quite widely used, has not been brought to bear in addressing the research questions.

The compressed timescale for completing this review has meant that less time has been available for consultation on the findings than in the case of previous ALRSG reviews. Instead of a conference where findings can be discussed by policy-makers and practitioners, consultation has been with the Review Group by email.

The low number of studies found that met the inclusion criteria for this review, with only seven providing evidence of high weight, meant that there is a limited evidence base for some conclusions, as the report makes clear in section 5.1.4.

5.3 Implications

5.3.1 Policy

Findings from the review suggest that policy-makers should note the following:

- Summative assessment by teachers has the potential for positive effects on students and on teachers, without the negative effects associated with external tests and examinations.

- Using teachers' assessment for summative purposes can support valid assessment of key learning processes as well as assessment of learning outcomes related to higher level cognitive skills.

- Summative assessment by teachers has the most benefit when teachers use evidence gathered over a period of time, and with appropriate flexibility in choice of tasks, rather than from an event taking place at a particular time. This enables information to be used formatively as well as summatively to adapt teaching.

- Using the results of student assessment for high stakes school accountability reduces the validity of the assessment, whether this is conducted by teachers or by external tests and examinations.

- Introducing new assessment practices can support beneficial change in teaching, providing that the techniques are well matched to learning goals and illustrate how students can be required to use important conceptual knowledge and leaning skills.

- Regulations for teachers' summative assessment should allow teachers opportunities to assimilate summative assessment into their practice and to design appropriate classroom programmes. When changes are made in assessment practices, time must be allowed for this assimilation to happen.

5.3.2 Practice

The following actions are likely to increase the benefit of teachers undertaking summative assessment of their own students:

- At all stages and for all purposes, students should be helped to understand the criteria by which their work is assessed. This is likely to mean providing and discussing examples that illustrate the practical meaning of the criteria.

- Teachers should make explicit to all concerned – colleagues, parents and students – the basis of the marks and grades they assign for internal school purposes. Achievement grades should not be influenced by non-academic factors, such as behaviour and participation, which should be reported separately as appropriate.

- When presenting assessment tasks to students, teachers should emphasise learning outcomes and not the attainment of a high grade, thus avoiding the encouragement of extrinsic motivation which leads to shallow learning.

- Teachers should internalise the progression in skills and understanding they aim to help students develop and interpret student performance in these terms rather than use checklists of specific unconnected behaviours. In this way, summative assessment helps teachers' understanding of learning goals as well as facilitates more detailed knowledge of their students.

- Schools should set aside time for teachers to discuss assessment issues, plan assessments and moderate their judgements of students' work. This not only improves the reliability of the assessment but enables teachers to use the process of summative assessment to help teaching and learning.

5.3.3 Research

The low number of studies found that met the inclusion criteria for this study, with only seven providing evidence of high weight, leads to an obvious implication that more research and more high quality research is needed in this area. Given the interest at high levels in government in making greater use of teachers' assessment in summative assessment, indicated the Primary National Strategy (DfES, 2003), the draft report of the Tomlinson Working Group on 14 – 19 Reform, and recommendations of the Daugherty report on assessment in Wales (Daugherty Assessment Review Group, 2004) there is some urgency in meeting the need for more research.

Particular research foci suggested by this review are:

- How teachers manage the dual roles as teacher and assessor
- The impact on students and on other uses of assessment of changing from tasks devised and marked externally to using teachers' judgements of students' performance in special tasks and in regular work
- The identification of factors that support teachers' use of summative assessment to improve students' learning experience; that is, how the formative use of assessment can be integrated with the summative use
- Direct comparison of different approaches used by teachers in summative assessment to investigate whether they make any difference to outcomes or to impact on students
- Investigating what information is actually used by teachers in their assessment and what impact this has on the curriculum experience by students
- The role of student self-assessment in summative assessment
- The impact on students of developing their awareness of success criteria and providing exemplification of learning goals
- What changes to accountability procedures would preserve the integrity of teachers' assessment and minimise pressures to give inflated grades or levels

6. REFERENCES

6.1 Studies included in map and synthesis

Abbott D, Broadfoot P, Croll P, Osborn M, Pollard A (1994) Some sink, some float: National Curriculum assessment and accountability. *British Educational Research Journal* **20**: 155–174.

Bennett SN, Wragg EC, Carre CG, Carter DSG (1992) A longitudinal study of primary teachers' perceived competence in, and concerns about, National Curriculum implementation. *Research Papers in Education* **7**: 53–78.

Bennett RE, Gottesman RL, Rock DA, Cerullo F (1993) Influence of behaviour, perceptions and gender on teachers' judgements of students' academic skill. *Journal of Educational Psychology* **85**: 347–356.

Brookhart SM, DeVoge JG (1999) Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education* **12**: 409–425.

Bullock K, Bishop KN, Martin S, Reid A (2002) Learning from coursework in English and geography. *Cambridge Journal of Education* **32**: 325–340.

Carter CR (1997/8) Assessment: shifting the responsibility. *Journal of Secondary Gifted Education* **9**: 68–75.

Cizek GJ, Fitzgerald SM, Rachor RE (1995/6) Teachers' assessment practices: preparation, isolation and the kitchen sink. *Educational Assessment* **3**: 159–179.

Flexer RJ, Cumbo K, Borko H, Mayfield V, Marion SF (1995) How 'messing about' with performance assessment in mathematics affects what happens in classrooms (Technical Report 396). Los Angeles, Centre for Research on Evaluation, Standards and Student Testing (CRESST). Available from: http://cresst96.cse.ucla.edu/Reports/TECH396.PDF

Gipps C, Clarke S (1998) *Monitoring consistency in teacher assessment and the impact of SCAA's guidance materials at Key Stages 1, 2, and 3*. Final Report. London: QCA.

Hall K, Webber B, Varley S, Young V, Dorman P (1997) A study of teacher assessment at Key Stage 1. *Cambridge Journal of Education,* **27**: 107–122.

Hall K, Harding A (2002) Level descriptions and teacher assessment in England: towards a community of assessment practice. *Educational Research* **44**: 1–15.

Hiebert E, Davinroy K (1993) Dilemmas and issues in implementing classroom-based assessment for literacy (Technical Report 365). Los Angeles, Centre for Research on Evaluation, Standards and Student Testing (CRESST). Available from: http://www.cse.ucla.edu/CRESST/Reports/TECH365.PDF

Hill M (2002) Focussing the teacher's gaze: primary teachers reconstructing assessment in self managing schools. *Educational Research for Policy and Practice* **1**: 113–125.

Iredale C (1990) Pupils' attitudes towards GASP (Graded Assessments in Science Project). *School Science Review* **72**: 133–137.

Johnston PH, Afflerbach P, Weiss PB (1993) Teachers' assessment of the teaching and learning of literacy. *Educational Assessment* **1**: 91–117.

Koretz D, Stecher BM, Klein S, McCaffery D (1994) The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues and Practice* **13**: 5-16.

McCallum B, McAlister S, Brown M, Gipps K (1993). Teacher assessment at Key Stage One. *Research Papers in Education* **8**: 305–328.

Morgan C (1996) The teacher as examiner: the case of mathematics coursework. *Assessment in Education* **3**: 353–375.

Pilcher JK (1994) The value-driven meaning of grades. *Educational Assessment* **2**: 69–88.

Stables A (1992) Speaking and listening at Key Stage 3: some problems of teacher assessment. *Educational Research* **34**: 107–115

Valencia SW, Au KH (1997) Portfolios across educational contexts: issues for evaluation, teacher development and system validity. *Educational Assessment* **4**: 1–35.

Whetton C, Sainsbury M, Hopkins S, Ashby J, Christophers U, Clarke J, Heath M, Jones G, Punchers J, Schagen I, Wilson J (1991) *A Report on Teacher Assessment.* London, SEAC.

Yung B (2002) Same assessment, different practice; professional consciousness as a determinant of teachers; practice in a school-based assessment scheme. *Assessment in Education* **9**: 97–117.

6.2 Other references used in the text of the report

Abbott D, Broadfoot P, Croll P, Osborn M, Pollard A (1994) Some sink, some float: National Curriculum assessment and accountability. *British Educational Research Journal* **20**: 155–174.

Assessment Reform Group (ARG) (2002) *Testing, Learning and Motivation*. Cambridge: Faculty of Education, University of Cambridge.

Barthes R (1975) *The Pleasure of the Text.* (trans. R Miller) New York: Hill and Wang.

Bell D (2003) Reporting England. Speech to the City of York Council's annual education conference. February. Available from: Office for Standards in Education

(OfSTED) News:
http://www.ofsted.gov.uk/news/index.cfm?fuseaction=news.details&id=1402

Black P (1993) Formative and summative assessment by teachers. *Studies in Science Education* **21**: 49–97

Black P (1988) *Testing: Friend or Foe?* London: Falmer Press

Black P, Wiliam D (1998) Assessment and classroom learning. *Assessment in Education* **5:** 7 –71.

Broadfoot P, Murphy R, Torrance H (eds) (1990) *Changing Educational Assessment: International Perspectives and Trends.* London: Routledge.

Broikou K (1992) Understanding classroom teachers' special education referral practices. Unpublished doctoral dissertation. Albany, NY: University at Albany, State University of New York.

Brookhart SM (1994) Teachers' grading: practice and theory. *Applied Measurement in Education* **7**: 279–301.

Brown A, Ash D, Rutherford M, Nakagawa K, Gordon A, Campione J (1993) Distributed expertise in the classroom. In: Salomon G (ed.) *Distributed Cognitions: Psychological and Educational Consideration.* New York: Cambridge University Press, pages 188–228.

Butler J (1995) Teachers judging standards in senior science subjects: fifteen years of the Queensland experiment. *Studies in Science Education* **26**: 135–157.

Carter CR (1997/8) Assessment: shifting the responsibility. *Journal of Secondary Gifted Education* **68**: 68–75.

Choi CC (1999) Public examinations in Hong Kong. *Assessment in Education* **6**: 405–418.

Crooks TJ (1988) The impact of classroom evaluation practices on students. *Review of Educational Research* **58**: 438–481.

Daugherty Assessment Review Group (2004) *Final Report. Learning Pathways through Statutory Assessment: Key stages 2 and 3*. Cardiff: Daugherty Review Group.

Department for Education and Skills (2003) *Excellence and Enjoyment – A Strategy for Primary Schools.* London: DfES.

Department of Education and Science (DES) (1987) *Task Group on Assessment and Testing (TGAT): A report.* London: DES and Welsh Office.

Donnelly JF, Buchan AS, Jenkins EW, Welford AG (1993) *Policy, Practice and Teachers' Professional Judgement: The Internal Assessment of Practical Work in GCSE Science.* Driffield: Nafferton Books.

EPPI-Centre (2002a) *Core Keywording Strategy: Data Collection for a Register of Educational Research. Version 0.9.7.* London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2002b) *Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research.* Version 0.9.7. London: EPPI-Centre, Social Science Research Unit

Flexer RJ, Cumbo K, Borko H, Mayfield V, Marion S (1995) *How 'messing about' with performance assessment in mathematics affects what happens in classrooms*. (Technical Report 396). Los Angeles: Centre for Research on Evaluation, Standards and Students Testing (CRESST).

Frederiksen J, Collins A (1989) A district's approach to educational testing. *Educational Researcher* **18**: 27–42.

Gilmore A (2002) Large-scale assessment and teachers' assessment capacity: learning opportunities for teachers in the National Education Monitoring Project in New Zealand. *Assessment in Education: Principles, Policy and Practice* **9**: 343-362.

Gipps C (1994) *Beyond Testing*. London: Falmer Press.

Gipps C, McCallum B, Brown M (1996) Models of teacher assessment among primary school teachers in England. *The Curriculum Journal* **7:** 167–183.

Gipps C, Brown M, McCallum B, McAlister S (1995) *Intuition or Evidence?* Buckingham: Open University Press.

Goldberg GL, Roswell BS (1999) From perception to practice: the impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment* **6**: 257–290.

Hall K, Webber B, Varley S, Young V, Dorman P (1997) A study of teacher assessment at Key Stage 1. *Cambridge Journal of Education* **27**: 107–122.

Harlen W (2004) A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes (EPPI–Centre Review). In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen W, Deakin Crick R (2002) A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI–Centre Review). In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen W, Deakin Crick R (2003a) A systematic review of the impact on Students and teachers of the use of ICT for assessment of creative and critical thinking skills (EPPI–Centre Review). In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen W, Deakin Crick R (2003b) Testing and motivation for learning. *Assessment in Education* **10**: 169–208.

James M (1998) *Using Assessment for School Improvement.* Oxford: Heinemann Educational.

Koretz D, Klein S, Shepard LA (1991) The effects of high-stakes testing on achievement: preliminary findings about generalization across tests. Paper presented at the Annual Meetings of the American Educational Research Association (Chicago, IL, April 3–7) and the National Council on Measurement in Education (Chicago, IL, April 4–6).

Koretz D, Stecher BM, Klein SP, McCaffrey D (1994) The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues and Practice* **13**: 5–16.

Lane S, Parke CS, Stone CA (2002) The impact of a state performance-based assessment and accountability program on mathematics instruction and students learning: evidence from survey data and school performance. *Educational Assessment* **8**: 279–315.

Linn R (1994) Performance assessment: policy promises and technical measurement standards. *Educational Researcher* **23**: 4–14.

Linn R (2000) Assessments and accountability. *Educational Researcher* **29**: 4–16.

Lubisi RC, Murphy RJL (2002) Assessment in South African schools. *Assessment in Education* **9**: 255–268.

McCallum B, McAlister S, Brown M, Gipps K (1993). Teacher assessment at Key Stage One. *Research Papers in Education* **8**: 305–328.

McMorris RF, Boothroyd RA (1993) Tests that teachers build: an analysis of classroom tests in science and mathematics. *Applied Measurement in Education* **6**: 321–341.

Massey A, Green S, Dexter T, Hamnett L (2003) *Comparability of National Tests over Time: Key Stage Test Standards between 1996 and 2001*. London: Qualifications and Curriculum Authority.

Maxwell G (1995) School-based assessment in Queensland. In: Collins C (ed.) *Curriculum Stocktake.* Canberra: Australian College of Education, pages 88–102.

Maxwell G (2004) Progressive assessment for learning and certification: some lessons from school-based assessment in Queensland. Paper presented at the Third Conference of the Association of Commonwealth Examination and Assessment Boards. Nadi, Fiji: March 8–12.

Messick S (1989) Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher* **18**: 5–11.

Morgan C (1996) The teacher as examiner: the case of mathematics coursework. *Assessment in Education* **3**: 353–375.

National Council on Education Standards and Testing (NCEST) (1992*) Raising Standards for American Education*. Washington, DC: NCEST.

National Research Council (NRC) (2001) *Knowing What Students Know. The Science and Design of Educational Assessment.* Washington, DC: National Academy Press.

National Union of Teachers (NUT) (1991) *Miss, the Rabbit Ate the 'Floating' Apple: The case against SATs: a report on the 1991 Key Stage 1 SATs*. London: NUT

Pollard A, Triggs P, Broadfoot P, Mcness E, Osborn M (2000) *What Pupils Say: Changing Policy and Practice in Primary Schools.* London: Continuum.

Popham WJ (2000) Modern Educational Measurement: Practical Guidelines for Educational Leaders. Needham, MA: Allyn and Bacon.

Radnor HA, Wilmut J, Burke P, Rainbow B, Parfitt G, Bellin W, Myhill D, Jennings S, Price N, Preece P, Baxter J and Skinner N (1995) *Evaluation of Key Stage 3 Assessment Arrangements for 1995. Final Report.* Exeter: University of Exeter.

Sadler R (1989) Formative assessment and the design of instructional systems. *Instructional Science* **18**: 119–144.

Shepard L (1990) Inflated test score gains: is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practice* **9**: 15–22.

Shorrocks D, Daniels S, Frobisher L, Nelson N, Waterson A, Bell J (1992) *Testing and Assessing 6 and 7 year-olds. The Evaluation of the 1992 Key Stage 1 National Curriculum Assessment*. UK: National Union of Teachers and Leeds University School of Education.

Torrance H (ed.) (1995) *Evaluating Authentic Assessment*. Buckingham: Open University Press.

Wood R (1991) *Assessment and Testing: A Survey of Research.* Cambridge: University Press.

Yung B (2002) Same assessment, different practice; professional consciousness as a determinant of teachers; practice in a school-based assessment scheme. *Assessment in Education* **9**: 97–117.

APPENDIX 1.1: Advisory Group membership

Review Group Membership

The members of the Review Group and their affiliations are as follows:

Members of ARG

Professor Paul Black, King's College, University of London
Professor Richard Daugherty, University of Wales, Aberystwyth
Dr. Kathryn Ecclestone, University of Exeter
Professor John Gardner, Queen's University, Belfast
Professor Wynne Harlen, University of Bristol
Dr Mary James, University of Cambridge
Dr Gordon Stobart, Institute of Education, University of London

Practitioners

Mr P Dudley Special Project Director, Classroom Learning, National College of School Leadership, and member of AAIA
Mr R Bevan, Deputy Head Teacher, King Edward VI Grammar School, Chelmsford
Ms P Rayner, Link Inspector for Primary Education, Nottinghamshire.

The ALRSG is advised by the following international experts:

Dr Steven Bakker, ETS International, The Netherlands
Dr Dennis Bartels, Director, President, TERC, Cambridge, MA. USA
Professor Lorrie Shepard, President, AERA, 1999–2000, University of Colorado
Professor Eva Baker, co-director of CRESST, University of California, USA
Dr T Crooks, Director, EARM, University of Otago, Dunedin, New Zealand
Professor Dylan Wiliam, Educational Testing Service

EPPI-Centre link staff

Dr David Gough, Deputy Director
Ms Zoe Garrett, Research Officer

APPENDIX 2.1: Inclusion and exclusion criteria

Inclusion criteria

Language of the report

Studies included were written in English. Although it was possible for translation from other European languages, the search strategy dealt with databases and journals in English and studies in other languages were not actively sought.

Types of assessment

Studies were included if they dealt with some form of summative assessment conducted by teachers. Studies reporting on purely formative assessment by teachers were not included, but those where the assessment was for both formative and summative purposes were included.

Study population and setting

Studies were included where they dealt with assessment procedures and instruments used by teachers for assessing pupils, aged 4 to 18, in school.

Study type and study design

Both naturally occurring and researcher-manipulated evaluation study types were considered to be relevant, as were designs including comparison of different approaches to summative assessment, surveys of conditions relating to the use of teachers' assessment for summative purposes and case studies of teachers' assessment used for these purposes.

Topic focus

Since teachers' assessment can be used in all subjects, studies from all curriculum areas were included. Studies were included both where evidence for the assessment was decided by teachers and judged against common criteria, and where assessment tasks or guidelines were prepared by others but the outcome was judged by the teachers.

Exclusion criteria

A: Not summative assessment. Studies were excluded if information was gathered for formative purposes only; aptitude tests and special needs assessment were also excluded.

B: Not assessment by teachers. Studies were excluded if they reported assessment of teachers or studies of school evaluation; also excluded were studies of teacher-administered tasks or portfolios that were graded externally.

C: Not related to education in school. This excluded studies relating to college students, higher education, nursing education or other vocational education.

D: Not reporting the impact of the process of assessment on students, teachers or the curriculum. Studies were excluded if the impact reported was a result of the outcome of the assessment and not the process.

E: Not research. Studies were excluded if they did not report empirical study of particular procedures of assessment by teachers; also excluded were handbooks and reviews and reports of instrument development or description, without a report of their use.

APPENDIX 2.2: Search strategy for electronic databases

Key terms:

Assessment by teachers	**Summative purpose**	**Relevance to school**	**Impact**
Teacher assessment Teacher-based asst Coursework asst Ongoing asst School-based asst Classroom asst Embedded asst Profile Portfolio Observation Process asst Moderation Grading	Summative assessment Examination Certification State/national assessment Baseline assessment Foundation assessment Transfer Transition Selection Graduation	School Infant school Primary school Elementary school Secondary school Community school Urban school Suburban school Private school State school High school Middle school Pre-school Kindergarten	Learning style Learning outcomes Achievement Teaching Teaching style Curriculum Students

These terms were entered as follows:

(teacher assessment OR teacher-based assessment OR coursework assessment OR ongoing assessment OR school-based assessment OR classroom assessment OR profile OR portfolio OR observation OR process assessment OR moderation OR grading) AND (summative assessment OR examination OR certification OR baseline assessment OR transfer OR transition OR Selection OR graduation) AND (school OR infant school OR elementary school OR secondary school OR high school OR community school OR urban school OR suburban school OR private school OR middle school OR pre-school OR kindergarten) AND (learning style OR learning outcome OR achievement OR teaching OR teaching style OR curriculum OR students)

APPENDIX 2.3: Journals handsearched

Journal search record

List of journals searched online (JOL) and by hand

Journal	Type of search	Dates searched	Number of articles found
American Educational Research Journal	Hand	1988–2003	0
American Journal of Evaluation	JOL	1988–2003	0
Assessment in Education	Hand	All 1994–2003	12
British Educational Research Journal	Hand	1985–2004	2
British Journal of Educational Psychology	JOL	1999–2004	2
British Journal of Educational Studies	JOL	1999–2004	0
British Journal of Educational Technology	JOL	1999–2003	0
Cambridge Journal of Education	Hand	1988–2004	2
Curriculum Journal	Hand	1988–2004	0
Educational Assessment	JOL and hand	1990–2004	13
Educational Evaluation and Policy Analysis	Hand	1988–2004	0
Educational Measurement	Hand	1993–2004	0
Educational and Psychological Measurement	Hand	1995–2004	1
Educational Research	Hand	1985–2003	5
Educational Research for Policy and Practice	JOL	2002–2004	1
Educational Researcher	Hand	1985–2004	1
Educational Review	Hand	1991–2003	3
Educational Studies	JOL and hand	1993–2004	1
Educational Studies in Mathematics	Hand	1996–2003	0
European Journal of Education	JOL	1999–2004	1
International Journal of Educational Research	JOL	1988–2003	1
International Review of Education	JOL	1999–2004	0
Journal of Curriculum Studies	Hand	1990–2004	0
Journal of Educational Measurement	JOL	1999–2004	1
Journal of Education Policy	Hand	1987–2004	0
Journal of Educational Psychology	Hand	1985–2004	2
Oxford Review of Education	Hand	1999–2004	0
Research Papers in Education	Hand	1984–2003	4
Studies in Educational Evaluation	Hand	1986–2004	4
Teachers College Record	Hand	1995–2004	1

APPENDIX 2.4: EPPI-Centre Keyword sheet, including review-specific keywords

V0.9.7 *Bibliographic details and/or unique identifier* ....

A1. Identification of report
Citation
Contact
Handsearch
Unknown
Electronic database
(Please specify.) ....

A2. Status
Published
In press
Unpublished

A3. Linked reports
Is this report linked to one or more other reports in such a way that they also report the same study?

Not linked
Linked (Please provide bibliographical details and/or unique identifier.)
....
....
....
....

A4. Language (Please specify.)
....

A5. In which country/countries was the study carried out? (Please specify.)
....
....
....

A6. What is/are the topic focus/foci of the study?
Assessment
Classroom management
Curriculum*
Equal opportunities
Methodology
Organisation and management
Policy
Teacher careers
Teaching and learning
Other (Please specify.)........................

A7. Curriculum
Art
Business studies
Citizenship
Cross-curricular
Design and technology
Environment
General
Geography
Hidden
History
ICT
Literacy – first language
Literacy further languages
Literature
Maths
Music
PSE
Physical education
Religious education
Science
Vocational
Other (Please specify.)........................

A8. Programme name (Please specify.)

....

A9. What is/are the population focus/foci of the study?
Learners
Senior management
Teaching staff
Non-teaching staff
Other education practitioners
Government
Local education authority officers
Parents
Governors
Other (Please specify.)...........................

A10. Age of learners (years)
0-4
5-10
11-16
17-20
21 and over

A11. Sex of learners
Female only
Male only
Mixed sex

A12. What is/are the educational setting(s) of the study?
Community centre
Correctional institution
Government department
Higher education institution
Home
Independent school
Local education authority
Nursery school
Post-compulsory education institution
Primary school
Pupil referral unit
Residential school
Secondary school
Special needs school
Workplace
Other educational setting (Please specify.)..

A13. Which type(s) of study does this report describe?
A. Description
B. Exploration of relationships
C. Evaluation
   a. naturally-occurring
   b. researcher-manipulated
D. Development of methodology
E. Review
   a. Systematic review
   b. Other review

Review-specific keywords

B 1 Object of impact reported (not mutually exclusive)
- B. 1.1 Students
- B. 1.2 Teachers/teaching
- B. 1.3 The curriculum

B. 2 Achievement assessed (not mutually exclusive)
- B .2.1 English (reading, writing, speaking, listening)
- B. 2.2 Mathematics
- B. 2.3 Science
- B. 2.4 Arts (music, drama, dance)
- B. 2.5 Other (specify)

B. 3.Origin of assessment task (not mutually exclusive)
- B. 3.1 Externally prescribed tasks
- B. 3.2 Selected by teacher from external bank or created using prescribed criteria
- B. 3.3 Set by teacher

B. 4 Type of assessment task (not mutually exclusive)
- B. 4.1 Special activity (timed or not timed)
- B. 4.2 Embedded assessment task
- B. 4.3 Portfolio for assessment
- B. 4.4 Project
- B. 4.5 Regular work
- B. 4.6 Other (specify)

B. 5 Use of result (not mutually exclusive)
- B. 5.1 Formative and summative
- B. 5.2 Internal (for grading, in-school records, reporting to parents)
- B. 5.3 External for accountability (high stakes for school)
- B. 5.4 External for certification (high stakes for student)
- B. 5.5 Other (specify)

APPENDIX 3.1: Mapping of the keyworded studies

EPPI-Centre Keywords

Table A3.1.1: Country in which the studies were carried out (N=23)*

Country	Number
Hong Kong	1
New Zealand	1
United Kingdom	12
United States of America	9

*mutually exclusive

Table A3.1.2: Population foci of the study (N=23)*

Focus	Number
Learners	23
Teaching staff	19
Parents	1

*Studies could be coded as having more than one population focus.

Table A3.1.3: Age of learners (years) (N=23)*

Age	Number
0-4	1
5-10	16
11-16	10
17-20	4

*Studies could be coded as having more than one age focus.

Table A3.1.4: Educational setting of the study (N=23)*

Setting	Number
Nursery school	1
Primary school	16
Residential school	1
Secondary school	11

*Studies could be coded as having more than one educational setting.

Table A3.1.5: Type(s) of study (N=23)*

Study type	Number
Description	4
Exploration of relationships	4
Evaluation: naturally occurring	12
Evaluation: researcher-manipulated	3

*Mutually exclusive

Table A3.1.6: Object of impact reported (N=23; not mutually exclusive)

Impact reported on	Number
Students	13
On teachers/teaching	17
On the curriculum	2

Table A3.1.7: Achievement assessed (N=23; not mutually exclusive)

Achievement assessed	Number
English (reading, writing, speaking, listening)	16
Mathematics	13
Science	10
Arts (music, drama, dance)	1
Other (specify)	4

Table A3.1.8: Origin of assessment task (N=23; not mutually exclusive)

Origin of assessment task	Number
Externally prescribed tasks	5
Selected by teacher from external bank or created using prescribed criteria	5
Set by teacher	16

Table A3.1.9: Use of result (N=23; not mutually exclusive)

Use of result	Number
Formative and summative	5
Internal (for grading, in-school records, reporting to parents)	15
External for accountability (high stakes for school)	3
External for certification (high stakes for student)	4
Other (specify)	3

Table A3.1.10: Type of assessment task (N=23; not mutually exclusive)

Type of assessment task	Number
Special activity (timed or not timed)	8
Embedded assessment task	5
Portfolio for assessment	3
Project	2
Regular work	15
Other (specify)	1

Table A3.1.11: Use of result for different types of assessment task (N=23; neither set of categories is mutually exclusive)

	Formative and summative	Internal (for grading, in-school records, reporting to parents)	External for accountability (high stakes for school)	External for certification (high stakes for student)	Research
Special activity (timed or not timed)	1	3	2	2	1
Embedded assessment task	2	5	0	0	1
Portfolio for assessment	1	1	1	1	0
Project	0	0	0	2	0
Regular work	3	12	2	1	1
Research	0	1	0	0	0

Table A3.1.12: Use of result for tasks of different origin (N=23; neither set of categories is mutually exclusive)

	Formative and summative	Internal (for grading, in-school records, reporting to parents)	External for accountability (high stakes for school)	External for certification (high stakes for student)	Research
Externally prescribed tasks	1	2	2	1	0
Selected by teacher from external bank or created using prescribed criteria	2	3	0	2	1
Set by teacher	2	12	2	1	2

Table A3.1.13: Object of impact reported for tasks of different origin (N=23; neither set of categories is mutually exclusive)

	Students	On the curriculum	On teachers/ Teaching
Externally prescribed tasks	4	0	4
Selected by teacher from external bank or created using prescribed criteria	2	0	4
Set by teacher	8	2	12

Table A3.1.14: Object of impact reported for tasks of different types (N=23; neither set of categories is mutually exclusive)

	Students	On the curriculum	On teachers/ teaching
Special activity (timed or not timed)	6	0	5
Embedded assessment task	3	0	4
Portfolio for assessment	1	2	2
Project	1	0	1
Regular work	7	1	13
Other	1	0	1

Table A3.1.15 Object of impact reported according to the use of the results (N=23; neither set of categories is mutually exclusive)

	Students	On the curriculum	On teachers/ teaching
Formative and summative	2	1	4
Internal (for grading, in-school records, reporting to parents)	7	1	11
External for accountability (high stakes for school)	2	1	3
External for certification (high stakes for student)	2	0	3
Other	3	0	2

Table A3.1.16: Object of impact according to educational setting of the study (N=23; neither set of categories is mutually exclusive)

	Students	On the curriculum	On teachers/ teaching
Nursery school	1	0	0
Primary school	7	2	14
Residential school	1	0	0
Secondary school	6	2	8

Table A3.1.17: Object of impact reported according to the achievement assessed (N=23; neither set of categories is mutually exclusive)

	Students	On the curriculum	On teachers/ teaching
English (reading, writing, speaking, listening)	7	2	13
Mathematics	6	1	11
Science	6	0	9
Arts (music, drama, dance)	1	0	1
Other	3	0	2

APPENDIX 4.1: Details of studies included in the in-depth review

Additional references in the appendices are not given in the reference list, but can be obtained from the original texts of the studies included in the review.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Abbott D *et al.* (1994)	Students Teachers/ teaching	Science	External for accountability (high stakes for school)	Evaluation: naturally occurring	Medium

Aims

To collect data to test the assumption that SAT assessment is reliable and investigate the question: 'Are SATs more reliable than teacher assessments made during the ordinary course of events?'

Study design

This was a case study of three Y2 classes in three different schools carrying out Science SAT. The study gives a detailed description of how the test was carried out and then, through interviews, finds out the teachers' reaction/feelings about the procedure.

Data collection

Observers 'used open-ended observation methods, making field notes and written records of teacher-child and child-child conversations and other interactions, while recording at intervals brief notes on such pre-selected categories as preparations for the SAT, arrangements for the rest of the class, extra help, if any, provided for the teacher concerned with the post-SAT events' (p 156).

Data analysis

No analysis as such. A straightforward account of the observation of the administration of SATs in three classes, in narrative and tabulated form.

Author findings

The results described here are those relating to the impact of the externally prescribed tasks – the SATs – on students and on the teachers. As the SATs had to be individually administered, this meant that the teacher could only give attention to one child at a time; other assistants were brought into the classroom to tend to other children. Teachers were uncomfortable about not being able to respond to other children when they were assessing one child. The teacher also noted their 'dislike of asking children to work alone, of being unable to offer help if asked without affecting the levels recorded' (p 167). Children were used to working in groups and found the individual situation unusual and may not have done their best work.

The impact on teachers was felt most strongly in relation to the use of time. Teachers thought the tasks were 'pretty pointless...Behaviour problems are being created... I just find the whole business time-wasting for the children and for myself. I could be teaching now' (p 166). They thought that they could obtain better information from observation in normal teaching interactions.

Author conclusions

They argue that the SAT activities are 'perfectly valid and useful lessons, but hardly useful means of assessment' (p 168). They felt some could have been abbreviated and were difficult to use. Some SATs took children so long to complete that they perceived that they provided no useful evidence of different levels of achievement. The unreliability would matter less 'if SATs were intended for internal consumption, as a guide to what has and what has not been successfully taught and learnt, rather than to produce league tables of schools'.

They also note that teachers are worried about the subjective judgements involved in SAT procedures ‘so it is arguable that Teacher Assessment is as trustworthy as SAT testing over most areas and can fulfil diagnostic and formative aims’. They discuss the pros and cons of supplementing TA with some form of standardised testing in limited areas in order to increase reliability (for summative purposes). The curriculum backwash would be likely to mean that teachers concentrated on what is tested. Moreover it ‘is hardly possible that assessment procedures in use with very young children can be standardised in any rigorous way as for GCE, for example’ (p 171).

In conclusion, ‘this study suggests that SAT assessment can never lead to reliable reporting of the comparative achievement of pupils or schools; in other words, to informing the market’.

Study	**Object of impact reported**	**Achievement assessed**	**Use of result**	**Study type**	**Overall weight of evidence**
Bennett *et al.* (1993)	Students	Cross-curricular skills and full range of subjects	Internal (for grading, in-school records, reporting to parents)	Exploration of relationships	Medium

Aims

To test a model relating tested achievement, gender, behaviour perceptions and teachers’ judgements of academic skill

Study design

794 students in Kindergarten, G1 and G2. Data were collected for correlational analysis to compute path coefficients to test out a path model of relationships between the variables.

Data collection

Behaviour perceptions: In Cleveland, behaviour grades for effort and conduct were given by the teacher; the mean of these was taken. In the Bronx, a single grade was given by the teacher.

Academic judgements: Ratings were made by the teacher in March and April, on a five-point scale (Grades 1 and 2 only) for mathematics, handwriting and reading comprehension, following common criteria. Means of these were used.

Grades given in June for report cards, across spelling, phonics, reading, mathematics, handwriting and English (not all for the Kg). Results of Einstein Assessment of school-related skills were obtained for each student.

Data analysis

The model hypothesised certain paths of influence. Standardised partial regression weights were computed for the model, using ordinary least-squares multiple regression in which each variable was regressed on its explanatory variables, beginning with the behaviour perceptions indicator and moving in sequence to the teacher academic judgements. Regressions were run separately for each grade within each district, thus permitting both location and grade to be treated as replications. (p 349)

Author findings

Gender differences: girls consistently obtained significantly higher behaviour grades than boys for Grades 1 and 2

Academic test scores and academic grades: no gender differences

Academic ratings gender differences (in favour of girls): only in Grade 1

From the path coefficients (standardised partial regression weights):

In Kg, behaviour grade consistently affected teachers’ academic judgements after controlling for gender, academic score and missing data. Large effect sizes.

Kg test scores consistently significantly affected academic grades (less so academic ratings).

Table 4 (p 351) shows correlations between gender and academic test performance. The table suggests the relationship is generally non-significant. Tables 5, 6 and 7 (pp 352 and 353) show path coefficients and multiple correlations for a hypothesised model for each grade. For kindergarten, it was found that the behaviour grade consistently affected teachers' academic judgements, whether they were grades or ratings.

Grades 1 and 2 had similar patterns to each other but differed in some respects from Kg. In all instances, gender was significantly related to behaviour grade, with effect sizes ranging from 0.23 to 0.37. Also, behaviour grade had a consistent direct effect on academic judgement (after control for gender, academic score and missing data). Academic test score showed a similar relationship with both academic grade and academic rating.

Indirect effects: only the path beginning with gender (through behaviour grade to academic judgement) had consistently direct effects. 'These indirect effects suggest that gender had a consistent effect on academic judgements that appear to have been mediated by teachers' perceptions of behaviour and that this effect was slightly stronger in the first grade than in the second grade' (p 350).

Author conclusions

Impact on students was for boys, predominantly, to be considered as less academically able than was the case. Thus they were given activities which did not stretch them and quite probably led to boredom and a vicious circle of poor behaviour. 'In all grades and in both districts, after controlling for tested academic skill and for gender, we found that teachers' perceptions of students' behaviour constituted a significant component of their academic judgements. In other words, students who were perceived as exhibiting bad behaviour were judged to be poorer academically than those who behaved satisfactorily, regardless of their scholastic skill and their gender. In Grades 1 and 2, however, boys were consistently seen as behaving less adequately than girls. As a result, teachers' perceptions of boys' academic skills were more negative than their perceptions of girls' capabilities' (p 351).

The magnitude was considerable. In Grades 1 and 2, a 1.0 SD change in behaviour grade produced about a 0.3 SD change in academic judgement; in Kg, the effect was closer to 0.4. By comparison, a 1.0SD change in tested academic skill produced a shift in academic appraisal for Grades 1 and 2 that was only marginally larger than that for behaviours; in Kg, this effect was essentially the same as that for behaviour.

The effects were 'surprisingly stable. With few exceptions, the results held across grades, school districts and outcome criteria' (p 351).

Conclusion: Behaviour perception is a potentially distorting influence.

Implications: 'First, these data reinforce the need to supplement teacher judgements with other objective evidence of academic performance when important decisions about students are made'. Second, 'the need for more concerted effort toward making teachers aware of the potential influence of student behaviour on their academic appraisals' (p 353).

Study	**Object of impact reported**	**Achievement assessed**	**Use of result**	**Study type**	**Overall weight of evidence**
Bennett *et al.* (1992)	Students Teachers/ teaching	English (reading, writing, speaking, listening) Mathematics Science	External for accountability (high stakes for school)	Evaluation: naturally occurring	Medium

Aims

To follow up an earlier questionnaire survey which aimed to 'investigate teachers' subject knowledge together with their self-perceived competence and professional skills required to teach effectively under a National Curriculum rubric; also to estimate the amount of inservice education and training (INSET) needed to support NC implementation as successive stages were put in place...' (p 54). 'As National Assessment and testing was applied formally and publicly to year-2 children at the end of key stage 1 during the summer of 1991, the questionnaire was revised to obtain information of teachers' experience with this process' (p 53).

Study design

A national questionnaire survey of primary teachers, with results expressed in terms of frequencies of different responses both fixed and open

Data collection

Self-completion questionnaire with separate questions for heads and Year 2 teachers. Although no examples of questions on assessment are given, the form these took is clear from the presentation of results.

Data analysis

Frequencies of responses to pre-set answers and categorisation of open responses

Author findings

Effects on teachers

'For TA the overall picture to emerge is that a large proportion of teachers felt that time was substantially reduced for normal classwork, and that although more teachers felt that were under additional stress (86%), class organization was largely unaltered (43%). A large proportion felt there was some value in this form of assessment for the planning of future lessons. There was some consciousness-raising and information learning regarding judging levels of attainment and acquiring knowledge about assessment techniques, but nearly half the teachers reported they did not feel professionally more competent as a function of their conducting TAs. For SATs most teachers felt time for normal classwork was even more reduced that for TAs and most registered additional stress' (p 68).

Effects on children

'Generally for a large proportion of the teachers (66%), discipline did not appear to be a great problem when TA was conducted. Discipline was, however, a problem for some of the time for 45% of teachers involved in the SATs. ...During TA most children enjoyed the extra attention, worked normally and were not upset or fretful, although a third of teachers signified that younger children were particularly disadvantaged for much of the time. The overall pattern, however, was one in which the equanimity of children was seemingly not unduly affected, and there were some positive gains for some as a result of the experience' (p 70).

Differences between SATs and TA

Major differences appear to relate to the processes of administration of administration and not in the type of information each revealed about students. One head was reported as noting: 'there are definite advantages to teacher assessment. It can be built into class planning, into normal class activities' (p 71).

Author conclusions

'The findings on assessment indicate considerable disruption to school routines and in classes, affecting particular groups of pupils during the main assessment periods. In 1991 there was class disruption from January–May' (p 77).

Researchers believe 'there is at least an imperative for some serious cost-benefit analyses to be undertaken before attempting to draw conclusions about implementation success regarding key-stage 1 assessment' (p 77).

There was inconclusive evidence as to whether TA and SATs gave the same information, but 'Notwithstanding, many primary heads, in their free-form responses, felt SATs were substantially a waste of time in their present form. What is clear is that inadequate resourcing of assessment in the form in which it is required to be implemented is unlikely to lead to optimal success' (p 77).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Brookhart and DeVoge (1999)	Students	English (reading, writing, speaking, listening	Research – theory building	Exploration of relationships	High

Aims

The purpose of this study was to test part of the theory of classroom assessment. A description of the level of perceived task characteristics, perceived self-efficacy, amount of invested mental effort, achievement, and the relations among these four events in two classroom environments was sought. The description would allow a test of the usefulness of the theoretical framework for interpreting results.

Study design

Four classroom assessment events in each of two classes were studied. To describe the classroom assessment environment from the point of view of an observer, language arts blocks of instructional time were observed on two different occasions. Language arts blocks consisted of reading, spelling, and language arts instruction; each of these used different texts, teachers taught them with different lesson plans, and students were made aware of the transitions between them (e.g., 'Get out your spelling books; it's time to do spelling.'). For each event, a pre-survey was administered to the whole class to collect evidence of perceived task characteristics (PTC) and perceived self-efficacy (PSE) to do the task. A post-survey was administered after the assessment but before students received feedback, to collect perceptions of amount of invested mental effort (AIME).

Data collection

The survey instruments were revised versions of pilot instruments tested with college freshmen (Brookhart, 1997a). Items from the pilot study instruments that covered the basic content of PTC (perceived task characteristics), PSE (perceived self-efficacy) and AIME (amount of invested mental effort), and had appropriate factor loadings on the pilot were rewritten in simple language for third Grade students. A five-point Likert scale was used, instead of the seven-point scale from the pilot and was illustrated with Snoopy pictures. Interview questions were written to prompt students to elaborate on the same concepts.

Data analysis

Descriptive statistics mainly were used (p 415). Interview data were coded into categories in the theoretical framework.

Author findings

Two generalisations can be made from the set of 32 interviews that seem to hold across classroom assessment events. First, previous assessment – that is, how they did on previous assignments – was the prime source of data students reported for basing their judgements about their self-efficacy to do current work. Second, students who talked about the specifics of their expectations for how they would do on the assessment reported particular examples of effort from which they took their information. Students who talked about general expectations for 'doing good' because 'I always do good' (p 421) did not report specific efforts at studying.

The interview data also illustrated the relations among instruction, perceived task characteristics, perceived self-efficacy and reported effort. These relations were evidenced in the way students expressed information from one category in another in their interviews.

The analysis of four classroom assessment events in each of two classes suggests three conclusions:

1. The model of the role of classroom assessment in student motivation and achievement (Brookhart, 1997a, 1997b; Figure 1) held in general: that is, there were relations among the

assessment task as students perceived it, their perceptions of their ability to do the task, their effort and their achievement. Both quantitative and qualitative data supported this conclusion.

2. Students' self-efficacy judgements about their abilities to do particular classroom assessments were based on previous experiences with similar kinds of classroom assessments. Results of previous spelling tests, for example, were offered as evidence about how students expected to do on the current spelling test. This finding is consistent with the model tested and also with self-efficacy research (Lepper, 1988; Schunk, 1994).
3. The relation between perceived self-efficacy and effort is not a simple one because students who perceived themselves to be so capable that the work would not be a challenge would not expend much effort. Specific prior experience with similar assessments may be necessary before students report investing less effort; that is, absent evidence that makes them sure they will do well. Students who perceive themselves as more efficacious will also tend to be students who report investing more mental effort in performance on an assessment. An exception might be for students who are performance oriented (Ames and Archer, 1988) and who thus might consider putting forth effort as an end in itself, by which they would be judged. Lack of variability among measures and small class sizes did not permit definitive conclusions about this relation' (p 422).

Author conclusions

'The theoretical importance of function significance feedback (Ryan, Connell and Deci, 1985) was illustrated in these results. Ryan and his colleagues called feedback informational if it gives students information they can use to see what they know and how they can do better next time. They call feedback controlling if it conveys judgement without information (e.g., good). However, they pointed out that students use similar feedback for different purposes; for example, one student might look at a test and simply see "B," whereas another might try to analyze which questions he or she got wrong. Informational feedback is seen as crucial to future learning. The results of this study suggest that judgemental feedback also may influence future learning through students use of it as evidence for their capability to succeed at a particular kind of assessment' (p 422).

The theoretical importance of goal orientations (Ames and Archer, 1988) was also demonstrated in the interview results. Students talked about mastery or performance-based reasons for investing effort in their work.

Teachers' explicit instruction and how they present and treat classroom assessment events affects the way students approach them When a teacher explicitly exhorts students to work 'to get a good grade', that teacher is on the one hand motivating students and on the other setting up a performance orientation that ultimately may decrease motivation. Teachers should make a point to exhort students to work for the satisfaction of learning or for its usefulness in accomplishing future work.' (pp 423-424) The authors question whether this is possible as long as the classroom assessment environment is affected by school district report card policies and the giving of grades.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Bullock *et al.* (2002)	Students	English and Geography GCSE coursework	External for certification (high stakes for student)	Evaluation: naturally occurring	High

Aims

To investigate:

1. The extent to which the original qualities attributed to coursework are achieved in current practice
2. The extent to which coursework contributes to the development of skills associated with independent learning, critical thinking and creativity
3. The influence of the demands of assessment upon students' learning

Study design

A series of interviews, with students (on two occasions) and with their teacher and parents (on one occasion)

Data collection

In-depth, semi-structured interviews took place with students (once in Year 10 and once in Year 11), parents, teachers and teacher-researchers. These aimed to gather perceptions of the links between coursework and creative learning.

A piece of English and Geography coursework from each of the students in the sample was selected for scrutiny by the team and evidence of creative learning was sought.

A half-day validation conference was held with the aim of validating and seeking to explain further, the emerging theories. The focused discussion constituted a final round of data collection and reporting within the project itself.

Data analysis

'The data were analysed using a mixture of conventional and electronic qualitative data analysis approaches.' No further details or information as to how the students' work was used (p 328).

Author findings

Perceptions of coursework

'All the pupils in the sample claimed some degree of independent learning as a result of their coursework. Those with a positive attitude to school (the great majority) found it motivating for the different skills that are practised and assessed through coursework tasks. There were perceptions of differences in the skills promoted by the two subjects areas. It was claimed that geography coursework tended to encourage literacy, numeracy, teamwork skills, and an ability to use initiative. On the other hand, it was suggested that English coursework encouraged insight, originality and imagination' (p 329).

There was good evidence that pupils perceive that they learn better through coursework (than through regular classwork). e.g. Student: 'When you're doing the research for coursework you find out a lot more than you would doing homework.' All respondent groups explained 'learned better' in terms of retaining skills and knowledge, finding out for themselves and being motivated. Reasons for this 'deep' learning were claimed to be the result of

- the assessment of coursework
- coursework requiring pupils to use different techniques for gathering and organising information
- the element of student choice

These claims were made more generally and strongly by the students and their parents than the teachers. However, parents stressed the challenging nature of the coursework and recognised the role of teachers in challenging pupils to provoke deeper thought processes.

'On balance pupils enjoyed coursework. Together with their parents and teachers, they valued the distinctiveness of this mode of learning and nature of the skills they acquired from it. ...In general, students were receptive to opportunities to learn more about learning – organising and managing their learning, how to become more independent – and demonstrated this in the interviews. Further benefits claimed were ownership of the project and freedom to organise their own learning and modes of working. The downside was the demands of time, deadlines, and the pressures from concurrent, similar work in other subjects' (p 330).

'Teachers, however, were less convinced that independent learning, creativity and critical thinking occurred in the majority of cases. ...teachers felt that, in general, other constraints, such as criteria for assessment, time and the need to achieve good grades prevented all pupils from reaching the optimum level of higher order thinking.'

Teachers were often reluctant to transfer the locus of coursework control entirely to students. They recognised the need for balance between prescriptive, directed approaches, leading to routine and repetitive task completion and presentation of findings, and freedom to develop original and creative approaches and ideas.

Parents indicated that, despite their being very willing to help, their children did not often want to involve them in their coursework.

The motivating impact of assessment

'Despite arguments that an authentic form of assessment such as coursework gives pupils the chance to demonstrate what they do know rather than test what they do not know, teachers thought their ideals of supporting creative learning were often limited in practice by the assessment framework emphasis on raising attainment and accountability. While it was clear the pupils were motivated by coursework, part of this motivation is related to the grade received at the end of the work. Just how much of the motivation could be attributed to extrinsic gains was difficult to measure' (p 336).

Students thought coursework good because it took the pressure off compared with exams...although coursework itself in not stress-free. Students recognised that what gained credit was the product rather than the process. In order to reach higher grades, most teachers 'now tend to coach students in specific techniques and look for evidence of these in their assessments. There can be no doubt that in a potential conflict between grades and independent learning, a higher grade would be preferred by students teachers and parents' (p 337). Thus the higher stakes tends to reduce the value of coursework for learning.

Students recognised that neither teachers nor students were compelled to follow particular coursework practices. This was despite claims about conflicts between managerial and accountability agendas and the forces of professional practices which influenced the majority of teachers' approaches. However in order to have access to higher grades, most teachers now tend to coach students in specific techniques and look for evidence of these in their assessments

Author conclusions

Students and parents agreed that a significant outcome from coursework was students' improved ability to initiate tasks and assume responsibility for their own work. This independence in turn engendered motivation for learning and positive feelings about the value of the different skills practised and assessed through coursework.

The assessment neglected the important processes that led to the products that were assessed. High stakes factors associated with accountability influenced the practice of completing coursework, to the extent where the promotion of higher order thinking was of secondary importance.

'Teachers assume that students will perceive the demands of learning and assessment in the same way that they do. In fact, despite teachers' assertions that marking schemes have been shared with students, the students tend not to understand what the assessment criteria actually require from them. Our research suggests that it is not sufficient to tell them; illustrations, examples and models are required' (p 338).

The authors also suggest that if there is a consensus to give value to transferable skills such as creativity, etc., a coherent inter-subject approach would support and develop these process skills efficiently and effectively.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Carter (1997/8)	Students	Mathematics	Internal (for grading, in-school records, reporting to parents)	Evaluation: researcher-manipulated	Low

Aims

To report on the implementation of Test Analysis, as alternative assessment method, used with gifted mathematics students, designed to shift the focus of assessment from the teacher to the students' (p 68).

Study design

The study collected information about the reactions of students as the process of Test Analysis was introduced and reports as narrative with examples and quotations.

Data collection

The study draws on test papers completed by the author's students and on those students' test analyses. There is no formal data collection as such.

Data analysis

Not stated

Author findings

'The Test Analysis creates several positive outcomes. Students actually took responsibility for their own learning.... It removed some of the pressure from nervous test takers. Since the test analysis could be very tedious and time consuming for the students, the number of mistakes, especially careless error, declined. Both teacher and student had less work if the student made fewer mistakes. The Test Analysis provided students with incentives to edit and check their work...students got enjoyment out of figuring out and analyzing their mistakes. The process gave the autonomous learner a chance to grow and feel good about the subject content. ..the focus shifted to constant learning rather than the teacher's judgement of a student's grade. On the negative side, the second part of the process – reading and assessing student comments – required a tremendous amount of work' (p 75). However, with a reduction in errors and familiarity, this declined. 'Students looked forward to the Test Analysis and protested if it was not used and they were denied the opportunity for improvement. 'Student assessment tended to be much more critical than my assessment' (p 71). Students were given approximately a week to work on the Test Analysis. They were encouraged to ask faculty and other students for help. The method was also effective when testing was done with a partner.

Author conclusions

'Providing students with personal feedback is very important to successful learning and teaching. The comments and analysis, although time-consuming, motivated all students. The opportunity was given for me to see the creativity, humour and skills that was sometimes stifled or overlooked in the classroom setting.

'The challenge of using alternative assessment is to accurately measure the amount of mathematics that students have mastered. For some students traditional assessment techniques are stumbling blocks to success. The goal of Test Analysis is to provide every student with multiple avenues of expression to display the knowledge that they acquire while taking a course' (p 75).

Study	**Object of impact reported**	**Achievement assessed**	**Use of result (not mutually exclusive)**	**Study type**	**Overall weight of evidence**
Cizek *et al.* (1995/6)	Students Teachers/ teaching	English, Mathematics Arts (music, drama, dance), Science Other subjects taught in schools	Internal (for grading, in-school records, reporting to parents)	Exploration of relationships	Medium

Aims

'This study focuses on understanding current practices to help formulate constructive suggestions for improving the flow of achievement information to all concerned' (p162).

Study design

A survey was administered to 143 students at the beginning of introductory masters level course at a Mid-Western university. All students were teachers with a diverse range of experience and from both primary and secondary schools. The survey was administered during the first two weeks of the course to prevent influence from the course (course in measurement and evaluation required as part of continued certification).

Data collection

Survey covered background characteristics and assessment practices as follows:

- factors teachers considered when assigning grades
- what the final grade for a marking period represented
- sources of information teachers used in assigning grade
- frequency of major and minor assessment tasks
- source teachers used to get assessments
- number of total marks used when calculating final grade
- knowledge of other teachers' grading practices
- knowledge of relevant district policies

Data analysis

Data were analysed by practice setting (elementary, middle, high), by gender, by years of teaching experience. For each question, a chi-square test of independence was performed to investigate potential relations between responses and these background variables. Finally a logistical regression procedure was used. Teachers comments were recorded and classified.

Author findings

Frequency and course assessments: 75.2% gave minor assignments at least once a week; the remainder gave them less than one a week. 53.8% gave major tests about once every two weeks; the remainder gave them less frequently.

Factors considered in assigning grades: 83.8% considered percentage or number correct on the assignment, 51.5% considered student's ability, 43.3% considered class performance, 41.9% student's effort and 35.3% difficulty of assignment. With reference to giving final grades, most teachers used formal achievement measures, such as tests, quizzes, reports. More than half used other formal achievement-related measures (such as attendance and class participation), whereas 41.9% used informal measures (such as student's answers during class or their contribution to discussions); and 61% used informal, non-achievement-related measures (such as conduct).

Number of grades assigned and grading policies: on average, respondents used 24.3 grades per marking period when calculating students' final grades; more than half did not know how this compared with other teachers in their building, a slight majority indicated their district had a formal grading policy; 11.9% did not know if district did or did not have one.

Analysis of teacher comments: these give some insights into widely variable assessment practices and knowledge of both measurement fundamentals and district polices as found in quantitative analysis.

Comments not solicited but volunteered:

'Different teachers use different methods of measuring progress, but they measure the same behaviours, so, ultimately, the results are similar.' 'Teaching in a private school requires a more rigid grading policy. Our policy is homework. The policy is to collect it, grade it, and record it.' Most teachers recognised that there are many valuable education outcomes worthy of assessing and are willing to try to assess them all to give a final grade.

Although cognitive development is acknowledged as an important contributor to students' grades, it was mentioned only rarely and in passing. Comments from teachers seemed to emphasise the non-cognitive outcomes, 'getting the child through the level with a positive attitude and good memories is more important than a raw number grade'. Several teachers reported the practice of throwing out the worst test result per student and taking into account impressions of ability and effort, thus demonstrating strong 'success orientation' in teachers.

Author conclusions

'Perhaps the most revealing overall finding is our general inability to discover strong predictors of differential assessment practice... One of our most interesting finding was that, for teachers who have not had formal training in testing and grading, very few of the assessment practices we studied were found to be relating to years of experience in the profession' (p173).

Researchers were perplexed about variability and unpredictability of teachers' assessment practices as related to following findings:

1. Despite lack of training in assessment, most teachers develop their own tests, quizzes and examinations.
2. More beginning teachers developed their own assessments compared with experienced teachers, with this difference related to practice setting and gender.
3. The extent to which achievement-relating and non-achievement-related information is taken into account in the final grades differs by practice setting.
4. Despite the fact that nearly every school district has some kind of grading policy, only about one half of the teachers in this study said that they knew their district had a policy and few of these teachers were able to supply any details about their district's policies.

A large percentage of teachers use commercial sources as primary source for tests.

'One fairly consistent finding in the teachers' comments revealed what might be called a 'success bias'. With discernable regularity, teachers appeared to structure their assessment practices and combine formal and informal assessment information in ways that were most likely to result in a higher grade for their students' (p 175).

Researchers recommend that districts begin to consider, establish and disseminate information that would provide guidance to teachers about desirable assessments and grading practices. They also recommend that schools more actively pursue engendering cultures of collaborative practice, especially related to assessment.

Beyond this consistency, it is not at all clear that any interested group – administrators, teachers, parents or even students and teachers themselves – can confidently glean the meaning of the grades students receive.

The finding that teachers were unaware of macro-level policy may not be surprising. However, in this study, teachers surveyed reported that they are also generally unaware of their colleagues' practices. Teachers interviewed for this study candidly admitted that they ignored district policies; several who acknowledged that they were unsure about what their colleagues did, vis à vis assessment and grading, also indicated that they preferred it this way (p 175).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Flexer *et al.* (1995)	Students Teachers/ teaching	Mathematics	Formative and summative Internal (for grading, in-school records, reporting to parents) Other; Research	Evaluation: researcher-manipulated	High

Aims

To study the effects of third grade teachers' work on performance assessment in mathematics on their beliefs and practices about curriculum, instruction and assessment and the ways they changed what they thought was important to teach, how they taught and how they assessed the performance of children (p 3).

Study design

Research carried out in school district with standardised testing programme in place and where they would be willing to waive standardised tests for two years in the schools taking part in the study.

Data collection

Research team collected data over one academic year (1992-93) from 14 third grade teachers in three schools (five in each of two schools and four in a third). Three interviews were conducted for each participant in the autumn, winter and summer. Interviews were taped and transcribed. Fifteen workshops from each school were read and coded. For the second round of analyses, six workshops were selected from each school which particularly addressed project goals.

Data analysis

Stage 1: All five authors read the same two transcripts (one interview and one workshop) to develop a tentative coding scheme. Scheme then went through two more iterations and before the final coding scheme listed in Table 1 (p 11).

Stage 2: 'Cases' were developed of each of the six targeted teachers (i.e. summaries of data organised according to several key areas).

Final stage: This entailed looking across these cases for themes that best describe effects of intervention; this final analysis addressed the initial research questions.

Author findings

Researchers found themes emerged in three key areas of analysis:

a. Beliefs and practices about how children learn: two main areas

(i) Differences among children

Most teachers believed some children were more capable of doing mathematics than others. As the year progressed, some teachers were surprised at how much third graders could do and became more willing to increase their expectations. By spring, most had a view of the developmental continuum for third graders that included higher order thinking.

(ii) Teaching children in small steps and keeping them comfortable

Most teachers believed that children learn mathematics by having mathematical concepts and procedures explained in small steps. For several, teaching students to do computations without understanding was also acceptable, because doing something successfully that others could do would raise self-esteem. All believed that experiential learning has some place though this varied at start of project. Even at the end of Year 2, teachers were concerned that children needed careful guiding, although some teachers wanted to set challenging problems.

b. Beliefs and practice about what is important to teach in school mathematics: two themes

(i) Computation

All teachers talked about importance of knowing and understanding facts, skills and computation. In the autumn, computation valued predominantly. This remained the primary focus but view of 'understanding' a process broadened for most teachers by end of year.

(ii) Problem solving and explanations: As the year progressed teachers gave more importance to strategies for problem solving, being able to explain how problems are solved and how procedures are done. Teachers expected excellent students to catch on quickly and included this in their description of 'excellent student'.

c. Instruction: three themes

(i) Shift in instructional practice

Shift during year toward using manipulatives, hands-on small group activities, problem solving and explanations. All teachers adopted Marilyn Burns multiplication replacement unit which has this focus.

One teacher reported the project had made her see 'how you change your instruction so that you're making children think more, more engaged, relating it to their everyday life'.

(ii) Student learning

Teachers reported they thought children were learning more and had better understanding. By the end of the year, students could solve problems and give explanations at a level that surprised many teachers.

(iii) Difficulties with new instruction

Some teachers had difficulties with the content and organisation of instruction alternative to the text. Although they agreed students had a better understanding, they were concerned with their knowledge of facts and appropriate skills. Also, leaving the textbook placed added burden of organising new materials. Many found this overwhelming and this was remedied by organising release time.

Assessment

A set of themes emerged: (a) by the end of the year, teachers were using more authentic evidence to assess what students know; (b) in the spring, teachers reported knowing more about what their students know; and (c) teachers encountered many difficulties with performance assessment.

Teachers' knowledge of students: teachers knew more about students from performance assessments. Most claimed performance assessment gave them new and deeper insights into children's thinking and understanding.

Author conclusions

Summary: The effects of the first year of the project on teachers' practice on instruction and assessment were numerous. Teachers were using more hands-on activities, problem solving and explanations. They were also trying to use more systematic observations for assessment. All agreed students had learned more and that they knew more about what their students knew. All struggled with revised instruction and new assessments. Many felt overwhelmed but got generally positive feedback from their own classes: that is, children had better conceptual understanding, could solve problems better and explain solutions. The teachers' response was to attempt further change in assessment and instruction practices and become more convinced of benefit of such changes.

Researchers found changes in assessment and instruction were mutually reinforcing. By the end of the year, many were using activities more closely aligned with National Council of Teachers of Mathematics (NCTM) standards, as was intended by the project.

The introduction of performance assessment provided teachers with richer instructional goals than mere computation and raised their expectations of what children could do. Teachers had concern for the comfort of their students and awareness that solving problems made students initially less comfortable than learning and performing computations.

They found teachers also felt uncomfortable with some of the changes until these were incorporated into their belief systems. 'While we did not try to change beliefs directly, we know we affected beliefs through changes in practice. There is no doubt that changes in beliefs alter practice, but it is also the case that shifts in practice may lead to shifts in belief, which can, in turn, further affect practice.... in other words changes in beliefs and changes in practice appear to be mutually reinforcing' (p 34).

'…what is abundantly clear is that the change that occurred did not result from anything we told teachers to do, but form their experiences with the ways performance assessments improved their classrooms' (p 35).

They found some teachers who didn't want to engage in the project, believing about performance assessment that 'this too shall pass'.

'It's about a lot of slow often painful, hard work for both teachers and staff developers. It's about the delight when the teacher who argues most vigorously about the changes says "I've changed my instruction. I mean I have to; I mean if I'm going to assess kids differently, I have to teach differently"'.

Researchers conclude that their work gives teachers confidence, that with continuing support, teachers are making even more changes. Question of persistence or stability cannot be answered in real time.

Change occurred from teachers' experiences with the ways performance assessments improved their classrooms and not from what researchers told them to do.

Teachers need a lot of support for changes they are expected to make and need permission to go slowly and make changes over a period of time. They also need many chances to try things out with children, and help in discussing and interpreting their classroom experiences.

Teachers need a lot of encouragement for all extra time and hard work. Staff developers must expect stops, starts and backward motions, and realise that not all teachers are at the same start point; not all interventions will work for all teachers; and each teacher will adopt different changes according to their existing beliefs and practices.

Results are not a clean sweep: It is not a matter of 'show the tasks and the teachers will use them' nor is it a matter of 'have teachers use performance assessment, and they will change their instruction'.

The researchers are not making an argument for high stakes enforcement of externally mandated performance assessment. It is not about forcing; rather it is about a lot of slow, often painful, hard work for both teachers and staff developers.

Study	**Object of impact reported**	**Achievement assessed**	**Use of result (not mutually exclusive)**	**Study type**	**Overall weight of evidence**
Gipps and Clarke (1998)	Teachers/ teaching	English (reading, writing, speaking, listening), Mathematics Science	Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

'Our research brief was to carry out a small scale programme of work to evaluate the use being made of SCAA's consistency materials, to provide an analysis of teachers' perceptions of Teacher Assessment, and to outline the nature of any further support that the SCAA should provide in the future' (p 17).

Study design

Data were collected through questionnaires sent to a large stratified sample of schools identified by SCAA. Three hundred were sent and a return of 150 was hoped for. Questionnaires were devised to cover all issues pertinent to this project and sent to assessment coordinators and/or secondary heads of English, mathematics and science departments. A shorter version, without questions concerning whole school policy, was sent to Year 2 and Year 6 teachers. Each type of school was sent the appropriate number of questionnaires. Interviews were carried out in 24 case study schools chosen as a purposive sample.

Data collection

Questionnaires were sent to assessment coordinators and/or head of English, Mathematics, Science, as appropriate, for specific school.

Interviews were conducted face to face and lasted between 30 minutes and one hour.

Data analysis

Questionnaires were largely multi-choice with some open-ended questions. Closed and multi-choice were processed using SPSS. Cross-tabulations were provided by age range of school and role of teachers. Open-ended responses were processed using standard qualitative analysis.

Author findings

1. *SCAA Consistency publications*

   66.6% said 'Guidance for Schools' was used to inform or develop school policy on consistency in TA. 30% had not used the publication. The booklets were used most often during standardisation meetings and at the end of key stage for assisting TA judgements. 85.4% said materials have helped ensure consistency of TA levels. More teachers from primary schools and heads of English believe this. Interviews found that more secondary teaches have not used materials for this purpose.

2. *Teacher assessment consistency issues*

   All heads of English described standardisations meetings and informal discussions about students' work as main strategies for ensuring consistency. 7 of 12 developed departmental portfolio. 9 of 12 heads of Mathematics said standardisation was meeting main strategy. Five had departmental portfolio. 7 of 11 heads of science described standardisations and production of departmental portfolios as main strategies. Key Stage 1, 2 and 3 assessment coordinators agreed that standardisation meetings and departmental portfolios were the main strategies.

   Standardisation meetings: 42% of primary schools have meetings more than once a term. Most secondary schools (over 75%) have them once a term or less. 98.8% of all teachers who said they had whole school or department meetings found them to be effective. All teachers who said they had meetings with smaller groups of teachers found them to be effective.

3. *General teacher assessment issues*

   All teachers, primary and secondary, said they did not have enough time to fulfil their role as an assessment coordinator and used evenings to carry out some of their work. Assessment coordinators asked whether they had assessment policies and whether these contained changes since the Dearing Report. Most secondary assessment coordinators said school policy did not contain changes as it was written as a set of general principles. For English departments, 13 of 14 had changed by abandoning tick lists and using level descriptions at the end of a Key Stage. 7 or 12 mathematics departments had changed and were now using level descriptions at the end of each Key Stage. All 11 science departments had changed policy by using level descriptions at the end of the Key Stage. All primary schools had changed policy; four said planning was now incorporated into assessment; four used 'best fit' instead of tick charts; and two now tracked significant achievement only.

   End of Key Stage TA levels: schools were asked how long it takes them to assign TA levels. There was a varied response, but, overall, these were much quicker for secondary schools and took from three hours to three weeks for primary schools.

   All teachers said test levels had not influenced TA levels as TA levels were completed before the tests; the exception were three heads of English and one head of science.

   All teachers were asked whether TA subject levels should be reported alongside Attainment Target levels: 55.9% said both should be reported; 31.5% said only subject levels; and 12.2% said only Attainment Targets, while 17.3% did not answer.

4. *Level descriptions*

   All were asked if they welcomed the introduction of level descriptions: 82.4% said they welcomed it. Interviewed teachers were asked about their opinions of an eight-level scale; responses varied, but a clear theme emerged around the scale providing a continuum; it was more manageable than the previous system.103 teachers said 'best fit' approach works when backed up by agreement trialling. Mathematics and science teachers tended to use more precise ways of grading pupils. It was found that primary teachers and heads of English departments were more likely to use best fit judgements in relation to children's portfolios.

5. *Further teacher assessment support*

   Teachers were asked what support they needed from SCAA for core and non-core subjects.

   English: 64.7% heads of department; 56.7% Y2 teachers; 58.7% Y6 teachers; 44% Key Stage 1 and 2 AC would like more exemplification documents. Need is slightly greater in Key Stage 3. Mathematics: 4.9% heads of department; 55% Y2; 58.7% Y6, 67.7% KS 1 and 2 would like more exemplification, so the need is greater in KS 1 and KS 2. Science: 64% heads of department; 50% Y2; 65.2% Y6 and 67.7% KS 1 and 2 would like more, so slightly more need in KS 2 and KS 3.

   Other: 10 teachers made comments and most common was 'would like to be left alone to teach'.

   With reference to INSET, it was found that KS 3 assessment coordinators would like INSET for non-core subjects more than primary teachers.

   With reference to cross-school agreement trialling, many primary school teachers had been involved but very few secondary school teachers. There was very little cross-phase trialling provided to schools by LEAs.

   With reference to INSET, this was provided by LEAs mainly for primary schools; up to half got supply cover. There was a greater need for more in-school INSET about end of Key Stage levelling than other types of training, followed by LEA INSET as the next most popular option. There was a greater need for in-school INSET to support process of ongoing assessment and record keeping than other types of training. In-school agreement trialling was seen as a training need by 69.6% of Key Stage 3 assessment. Teachers are possibly more aware of the issues involved in achieving consistency than their head of department colleagues.

   Interviewed assessment coordinators were also asked for their views on training needs for the end of key stage levelling. Their responses supported the questionnaire findings.

Author conclusions

With reference to TA issues, 'it seems that more interviewed teachers believe that consistency of level judgements for Teachers Assessment is not possible, mainly because of the inevitable differences in interpretations of the assessment criteria. Those who believe it is possible see agreement trialling as the only way to achieve it. Teacher Assessment is clearly important, for various reasons, but, like tests, will never be completely "accurate". Consistency is likely to be possible with regular ongoing trialling' (p 56).

With reference to the end of Key Stage Teacher Assessment levels, most heads of department found it manageable and most primary teachers did not; the difference is clearly the style of marking and testing between primary and secondary schools which causes this difference in manageability.

With reference to setting, it was found that most mathematics and science departments appear to use both TA and test levels, whereas most English departments either do not set, or mainly use, TA levels. Perhaps this is inevitable, given the subjective nature of the grading of English tests.

The following themes emerged with reference to future role of TA:

- TA is vital; it is the core of the learning process.
- TA judgements need to have higher status than they have at present.
- TA informs future work, raises children's awareness and raises the profile of achievement.
- TA is essential because children do not always perform well at tests.
- TA is more important than external tests because it is ongoing and is not focused on narrow criteria.

A variety of sources of information is used to decide levels, written work, classroom tests and assessment activities, observations, dialogue, homework (secondary schools) and memory (primary schools). Y6 teachers were more likely to move towards more formal sources of information, such as marking and ongoing tests than Y2 teachers.

Final comments by teachers: three main themes emerged with reference to Teacher Assessment:

- Any assessment structures must be able to fit into the time available and/or supply cover provided.

- Assessment demands get in the way of teaching.
- TA is more useful than tests but it has too low a status, because test results dominate the league tables.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Hall and Harding (2002)	Teachers/ teaching	English (reading, writing, speaking, listening) Mathematics Science	Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

To assess the extent to which a community of assessment practice is evident in schools in relation to the use of level descriptions (LDs). By ‘assessment community’ is meant a shared understanding among staff of the goals of National Curriculum assessment in general and TA in particular; a shared set of processes for the pursuit of these goals; and a common usage of a range of tools like the LDs, portfolios and exemplification materials to help staff with their assessment tasks.

Study design

Six schools were selected for participation. Interviews were held with all Year 2 teachers and the assessment co-ordinators in two years. Year 3 teachers were interviewed in the second year. One assessment meeting was observed in one school. LEA assessment advisers were interviewed in both years.

Data collection

Interviews of LEA advisers of assessment and of teachers of 7-year-olds and assessment coordinators in all six schools in two consecutive years (1998 and 1999). Observation of one assessment meeting in one of the schools. Collection of documentary evidence, such as portfolios, record sheets and school and LEA assessment documents. All interviews were audio recorded and later transcribed in full.

Data analysis

Qualitative analysis following the procedures of Miles and Huberman (1994): ‘…on the production of typed transcripts and field notes, we individually scrutinised the evidence in relation to the research focus, themes from the literature and patterns that seemed to be emerging. We made written notes on the scripts, highlighting what we perceived to be the key themes and continuities and contradictions in the data. Our thinking and interpretations were refined as we revisited and reread the different transcripts ...and as we clustered and categorised the evidence, following repeated checks, matching and cross-checking, especially cross-referencing the data bases for the two years and for the six schools....The second year of data provided a useful means of validating the evidence made available in the first year, so themes emerging from the analysis of the first year were revisited and further probed in these interviews’ (p 4).

Author findings

Overall the authors identified two conceptually different approaches to TA at school level, which they called ‘collaborative’ and ‘individualistic’. The former exhibited many of the characteristics of ‘an assessment community’, whereas teachers in the latter tended to work largely in isolation from their colleagues. Key elements of assessment identified with these positions were goals, tools and processes, personnel and value system.

Differences in these key elements were tabulated (p 6). In brief, collaborative schools showed compliance and acceptance of goals (contrasted with reluctant compliance and resistance); sharing of interpretation of LDs, active portfolios, planned collection of evidence, common language (contrasted with little sharing of interpretations of LDs, dormant portfolios, evidence not much used, assessment often bolted on, confusion about terms); whole school involvement and aspirations to involve parents and students (contrasted with Y2 teachers working as individuals and no grasp of the potential of enlarging the assessment community).

Assessment was seen as useful, necessary and integral to teaching (contrasted with assessment seen as imposed and not meaningful at the level of the class teacher). Interviews with the LEA advisers showed that they had built up a considerable expertise in TA, use of portfolios and some formative use of the assessment information. However, interviews with teachers showed that they had limited access to this expertise: for a variety of reasons, some relating to the large number of initiatives which resulted in TA being put 'on the back burner' (p 6). 'In such circumstances, teachers are left to depend on one another for support. In four of out six schools, TA was presented as the business of the whole staff and individual efforts were supported and bolstered up by a collective machinery that involved discussion and decisions on the key elements... However, with the exception of one school, all such collaboration occurred at the end of the teaching day and increasingly competed for time with a host of other initiatives.' In the other two schools, teachers perceived themselves as working alone, so the process of assessment had not led to more teacher-teacher discussions about children's learning.

'Although practices designed to enhance consistency and validity of assessment decisions were in place in four of our sample schools, we detected a decline overall in the level of collaboration in the second year of data gathering. Teachers were becoming more preoccupied with the National Literacy strategy....' (p 8).

In relation to parents, all teachers cast themselves in the role of information givers and, to varying degrees, as interpreters of TA terminology. Schools where teacher met among themselves to discuss TA went to greater lengths to make the results and processes meaningful to parents.

Differences between the schools were even sharper in relation to involving students. In two schools, there was little appreciation of the potential of using samples of their own (or other students' work) with students to help them get to grips with success criteria. In only a few of the schools was there any real grasp of the importance of involving students in the assessment process.

In two schools, all teachers interviewed 'expressed scepticism about the pedagogical usefulness of the results. Typically responses in both these schools to a question about the potential of the results to inform planning and teaching referred to 'lack of time', being 'too busy', the grades or levels being 'vague and lacking in differentiated information' and they themselves feeling overwhelmed by 'doing assessment' (p 11). In the circumstances of other pressures, teachers concentrated on 'getting it done' rather than thinking about implications for the child.

Author conclusions

Whilst there is evidence of the emergence of an assessment community of practice within some schools, such communities are confined mainly to the teachers within those schools.

The potential for both learners themselves and their parents to be more actively involved has not been fully explored and exploited.

The extent to which the process is having a positive impact on teachers' planning, etc., is limited by the time allocated to it and the competing pressure of other centrally imposed initiatives. 'The fact that funding was not made available for teachers to moderate the TA results served to tell teachers that the results of the external testing programme were prioritised over TA....The fact that TA, more than most other recent initiatives introduced into schools, depends on teachers exercising their professional judgement meant that teacher professionalism was enhanced and affirmed accordingly. Its diminished status, therefore, threatens that sense of professionalism' (p 12) .

'The problem in recent years is that teachers' learning has been more technical than professional, focused on the short-term implementation of Government priorities. The lack of support for resources to support teachers in their interpretations of LD's and their application of TA is not surprising ...since the assumption is that the solution is technical and not about professional learning' (p 13).

The more assessment is presented as a technical rather than a professional matter, as is necessary for TA to be used dependably, the more difficult it will be to reach the situation where teachers' judgements can be trusted.

The authors argue that the quality of teaching and learning inside the classroom is strongly influenced by the quality of the professional relationships teachers have with their colleagues outside the classroom, so that there was potential for increasing quality through building professional cultures among primary teachers in the wake of the NCA. Now these cultures are no longer supported and the ground gained earlier could be receding.

The authors conclude that the potential for effective TA is not being taken up. It is not being used effectively to inform future planning and in many schools it is unreliable because of the lack of attention given to moderation and use of exemplification materials. TA is becoming unreliable in the face of pressures to raise standards and 'the ground that was gained (in quality teacher assessment) in the early and mid 90s could be receding' (p 13).

The authors conclude that the quality of TA is in decline as a result of the removal of funded opportunities for moderation across schools and because of the increased attention being given to other priorities (literacy and numeracy strategies) at the same time as requirements to publish TA data were removed.

They conclude that what they see as important 'communities of assessment practice' are vulnerable and fragile. This, combined with the pressure to raise standards, is putting some individuals under pressure to assess at higher levels than they would have done. Where individuals are working alone, there are fewer checks and balances to prevent this. One teacher is quoted as saying that 'Up to now our TA has been "If you're not sure (about a higher level), then don't give it" and now you're having to think "Well if you're not sure, we'd better give it."'(p 12).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Hall *et al.* (1997)	Students Teachers/ teaching	English Mathematics Science	Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

The study aims to review approaches to TA 'at the level of school policy and at the level of classroom practice' (p 108). The purpose is to document ways schools respond to the requirement of TA. Specifically, the study sought information on:

1. the procedures used in TA
2. the manageability of the TA process
3. the impact of this process on children's learning
4. the impact of this process on teachers' pedagogical approaches
5. the management of moderation

'Also sought and arising from the above are teachers' own understandings of the purposes of TA, their perceptions of the accountability dimension of TA and the extent to which TA practices are influenced by school policy. The aim of this exploratory study is to describe and explain rather than to predict and generalise' (p 108).

Study design

Teachers in 45 primary schools were interviewed in June/July after completing their KS 1 TA

1. Sampling of Year 2 teachers in a sample of schools from a single LEA somewhere in England or Wales to ensure diversity of schools with regard to specified variables
2. Data collection: (a) semi-structured interviews held mainly at schools, lasting between 30 and 90 minutes (sometimes >1 teacher interviewed at a time) and (b) documentary evidence in the form of policy statements and samples of children's work and reports

3. Data analysis: (a) quantitative data on frequency of reports of use of different TA methods aggregated and placed in tables and (b) qualitative analysis of teachers' descriptions of their approaches and perspectives

Data collection

One-to-one interviews arranged: through 'phone contact; format: semi-structured, no detail of recording/note-taking

Content: 'Focused on specific aspects of TA, teachers were given opportunities to elaborate on any facet... they thought relevant.... allowed teachers to reveal their own attitudes to and understandings of TA and the strategies they use to assess their pupils' (p 108).

Interviews were mainly held at schools, lasting between 30 and 90 minutes. Sometimes more than one teacher was interviewed at a time.

Policy statements and samples of children's work were also mentioned, but no detail given.

Data analysis

Quantitative aggregation of data (frequencies of responses) and qualitative analysis which identified a 'model' of conducting TA which was the only one fitting the interview data.

A form of qualitative analysis appears to have been used to come up with: (a) a model that describes stages used for assessment over the second school year and to describe the main characteristics of each stage; and (b) other main themes arising from discussions.

The authors may be referring to grounded theory when they say that the first section of their results (see (a) above) 'presents an overview of results in the form of a grounded model derived from the empirical data' (Hall *et al.*, 1997, p 109).

Author findings

Teachers go through a series of stages in conducting TA, represented in the 'model'

1. assessment planning stage
2. observation stage
3. specific task stage
4. continuous review stage
5. levelling stage

The fourth stage is the longest, when the teacher not only gathers evidence fairly systematically but makes judgements about it. Stages 3 and 4 are recursive in that judgements made at stage 4 inform the allocation of tasks. A characteristic of stage 4 is that it is now largely, although not wholly, a formalised process of assessing the extent to which attainment targets have been attained' (p 111). This contrasts with the definition of assessment evidence at the second stage which is predominantly to do with making professional judgements on the broader aspects of development. The fifth stage refers to the allocation of a level to each child and occurs over a short period, four to six weeks before the end of the school year. The last two stages form a two-way process in that the levelling itself informs the updating of TA records and vice versa. Particular attention was paid to the assessment of process skills; this provided the greatest challenge to teachers. There was a concern about making fair and accurate assessment. It was in this area that teachers used 'intuition' rather more than systematic data and interpretation.

There was a sense of 'professional mistrust' amongst teachers (p 113). A wide range of assessment material was passed on from Year 1 teachers but these assessments by other teachers were treated with some caution, even suspicion.

A minority of teachers referred to assessing the broader aspect of children's learning (outside the NC requirements).

Impact on learning and teaching

Overall, the impact of TA on the quality of children's learning was perceived to be positive by the majority of teachers (63%) (p 119). The majority of teachers claimed that the main benefit to children's learning is the match which is facilitated between the experiences and activities provided and individual needs. This was especially emphasised in the case of students with SEN.

Teachers were unanimous in their claim that the need to assess caused them to plan in greater depth and to plan for the short, medium and long term. However, it also caused them to concentrate more on curriculum coverage rather than follow their own or children's inclinations and interests.

Concern expressed over issue of how and how often evidence of progress should be documented. Over 75% reported using record sheets and checklists.

Teachers also showed an awareness of the importance of regular, close study of children's work (e.g. annotated samples of work kept in portfolio). Manageability and effects of assessment on children's learning and teachers' practice were recognised.

Overall, the impact on children's learning was seen to be positive by the majority of teachers (63%). Teachers 'did not indicate that they were aware of or overly concerned with the accountability dimension of TA' (p 120).

'Teacher assessment gives you a better insight into children's ability and this makes you focus your teaching more. It is also beneficial in terms of continuity' (p 120).

The majority of negative comments related to manageability of the process and the burden resulting from the changing nature of assessment requirement (p 120).

It was also found that a result of national testing on core subjects meant that teachers devoted more time to these at the expense of a broader and more balanced curriculum (p 121).

Whole-school policies on assessment

There was a consistency between what teachers in a particular school said and what appeared in their school's policy.

36% of interviewees said whole-school policies were in place in their schools. A further 35% claimed they were in process of being formulated.

Author conclusions

The most significant aspect of the model is that TA is seen as an activity which influences all aspects of curriculum implementation: from curriculum planning before the school begins to summative, individualised reporting on each child at the end of the school year. In this sense it seems that attempts are made to integrate assessment into the act of teaching and not merely add it on to satisfy official requirements.

Teachers were adapting their practices in line with the assessment requirements and the consequences were enhanced learning opportunities.

However there were a number of negative aspects: for example, focus on a single year (Year 2) rather than the whole Key Stage.

The study identified the need for more research into the assessment of process skills; the effective use of ipsative and peer assessment; the balance teachers strike between assessment in the cognitive and the affective domains; and the extent to which teachers' practice in this classroom conforms to their own perceptions as revealed in this study.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Hiebert and Davinroy (1993)	Teachers/ teaching	English (reading, writing, speaking, listening)	Formative and summative Internal (for grading, in-school records, reporting to parents)	Evaluation: Researcher-manipulated	Low

Aims

An examination of the actual effects of introducing new forms of assessment at the classroom level. It examined the issues that arise over the course of staff development on classroom-based assessments.

Study design

A qualitative case study carried out across three schools. Schools were invited to participate and, if all third grade teachers were willing to take part, they met the research team and took part in seven meetings.

Data collection

Data were collected by tape-recording teacher meetings held in the schools and by taping interviews with teachers (although not used in this analysis). Field notes and transcriptions were analysed retrospectively over the course of the meetings to identify themes and dilemmas (obstacles to teachers' implementation of assessments) which were the focus of reporting and interpretation.

Data analysis

Categories were established from the transcripts and the data were then coded using these categories. Transcripts that represented the three schools and three periods of time – first two sessions, middle three, and final two – were read and notes were written down about the content of teachers' comments by two members of the research team. The category scheme was refined until transcripts could be accounted for by the category scheme. This category scheme was used to code the data for all of the transcripts.

Author findings

The results are given for each of the three schools in the study one at a time.

School 1

Encountered a dilemma about the amount of time that the assessments were taking from instruction. These teachers described lessons about students writing summaries and deliberated on appropriate ways to instruct and assess summaries. In discussing and developing a scoring rubric, teachers explored the issue of standards in scoring. Working out the scoring rubric enabled them to identify the elements in students' work that indicated 'good' work and ways of fostering it, such as getting students to score summaries.

School 2

There was evidence that using the assessment had led two teachers 'to adapt their instructional practices by initiating whole-class phonics lessons (p 14). While this strategy was questioned by other teachers, 'These statements suggest that the assessment had been the impetus for teachers' reflecting and acting on their instruction' (p 15).

School 3

'The topic of the conversations changed over the sessions - from defining literacy to developing a rubric that discourages formulaic responses to instructional strategies that foster students' thinking. Two characteristics remained constant across the sessions: the form of the conversation and continual references to the underlying processes that the assessment and instructional activities were intended to further' (p 17). Teachers recognised the limitations of their existing assessment in not giving them information about the quality of work which would inform their instruction. Teachers began to expect more of students once they themselves understood the progression in the underlying skills involved.

Author conclusions

Working on classroom-based assessments put demands on teachers (p 19).

It was difficult to embed the assessments into classroom instruction and they tended to be add-ons (p 20).

The assessments revealed difficult information about students' literacy levels. This presented an ethical issue about who to share this information with, especially to avoid labelling students (p 20).

The three schools had different definitions of literacy and assessment (p 20).

There was a tendency to focus on assessment techniques (in two schools) rather than on a school-wide effort to define the goals of literacy assessment (one school).

A three-month period is not sufficient to determine whether the implementation of assessments can change teachers' views of literacy and of literacy instruction (p 20). In many ways, the new assessments were seen as add-ons and were not firmly embedded in the underlying foundation (p 20).

The process of classroom-based assessment is not necessarily welcomed by teachers.

The authors quote Resnick and Resnick that it is not the testing per se that has narrowed the curriculum but the questions asked. Thus, changing the assessment tasks to mirror the critical processes of literacy should change the instruction. 'It is a rare assessment project that begins with a recognition of the transformation of goals and instructional practices in a state or district. Only one school spent time on a school-wide effort to define goals. Other teachers began with a focus on assessment techniques' (p 20).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Hill (2002)	Teachers/ teaching	English (reading, writing, speaking, listening) Mathematics	Formative and summative Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

'The purpose of the research ...was to describe how some New Zealand teachers dealt with the competing discourses described above (i.e. discourses emphasising the use of assessment for monitoring and accountability rather than the use of assessment to help learning) and the effect these discourses had on their assessment practices' (p 116).

Study design

An interview survey of 12 primary teachers in two case study schools. Interviews were conducted over two years, but it is not clear how many times each teacher was interviewed (reference to the 'first' interview indicates more than one but not stated). Interviews were transcribed and analysed and reported qualitatively.

Data collection

Interviews and class observation. No examples of questions or interview schedules are given.

Data analysis

Qualitative analysis using NUD*IST software.

Author findings

Most of the teachers in the study appeared to focus their attention on the ability of the students to achieve the objectives set for a specific unit of work. Typically, these teachers planned from the New Zealand curriculum documents, they mainly saw teaching and assessing as separate activities, describing instances of assessment as reasonably formal, planned events: that is, formative assessment was framed as 'teaching' not assessment. Teachers used extensive checklists, checking each child against achievement objectives. They did this because of perceived expectations of the school's policy. Their assessments were 'driven from above' (p 117).

Those who relied on their memories of what children could do said they focused on teaching not assessing. These teachers, however, about once a term set assessment tasks and checked off progress against the achievement objectives, again using a checklist system (p 119).

The impact on teachers' practice was framed mainly in relation to the accountability demands. 'Some teachers had managed to balance teaching and accountability demands while others appeared to have more difficulty dealing with these competing discourses' (p 120). Many focused almost entirely on checking progress against achievement objectives. 'The main drawback with most checking systems is that they are not geared to guide pupils in how to improve their learning nor can they diagnose individual strengths and learning needs' (p 120).

The evidence from the research reported here suggests that it is possible to use evidence collected for formative purposes for summarising achievement providing that the different uses are understood and that the teacher knows how to use the information in their teaching as well as how to report it for accountability purposes. However, the prominence of checklisting approaches for gathering evidence suggested that many teachers have found it very difficult to prioritise assessment for formative purposes following the reforms to education.

School-wide policies heavily influence teachers' practices.

Author conclusions

From these results it would seem that teachers would appreciate help to make connections between their own tacit 'craft' knowledge and the curriculum objectives, to re-establish the central place of formative assessment in learning and teaching, and to find non-intrusive ways of recording and reporting outcomes.

Rather than setting targets and reporting achievement against the achievement objectives in every curriculum area, schools could assist teachers to focus on priorities, such as literacy and numeracy. This could be done by developing policies that require information about only very important areas to be collected and reported, as well as use for improving teaching so that teachers can put their emphasis on using formative interactions in their teaching, rather than reporting on achievement targets.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Iredale (1990)	Students	Science	Formative and summative Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	High

Aims

'The purpose of the study was to investigate pupils' attitudes towards GASP' (p 133).

Study design

A cross-sectional study of the attitudes of first and second year secondary students' attitudes to the GASP scheme, using data collected by a questionnaire, interview (two students at a time) and students' essays. Questionnaire data were analysed quantitatively, exploring differences between genders, age and achievement levels, and essays and interview data used in illustration.

Data collection

Students essays: The students from three out of the six teaching groups in each year were asked to write about their opinions of GASP under the title 'What I like and what I don't like about GASP'. Students were encouraged to express their opinions honestly and did not add their names to their work. Most wrote between a quarter and a half-side of A4 paper.

The questionnaire: the questionnaire consisted of eleven items. It was completed anonymously by all first and second years during a science lesson (p 133).

Interviews: the interviews were 'semi-structured', based a pre-selected set of questions but with opportunities to follow-up related issues as they arose. The majority of the interviews involved two students as this proved to be most successful in eliciting detailed and thoughtful replies (p 134).

Data analysis

For the questionnaire, the number of students giving each of the three responses on each item was recorded and percentage responses calculated. The level of statistical significance of the results was tested using the chi-squared test and the 1% level set as the level of significance appropriate for drawing conclusions.

No information is given about how interview and essay data were analysed, although there is evidence in the presentation of results that they were categorised according to the attitudes they indicate.

Author findings

'The general impression gained from these results is one of enthusiasm and approval. A large majority of pupils (83%) believe that passing levels will help them in the future. At this early stage in their secondary school career some pupils appear to be looking ahead to GCSE and beyond. They see the value of the GASP scheme in providing a continuous-assessment route to GCSE and evidence of their achievements in science. 91% of pupils considered the scheme preferable to an end of year examination.

'GASP involves a comparatively high frequency of testing with each pupil taking on average one test every four weeks. The majority of pupils (62%) appear content with this level of testing however, only 15% feeling that it is too frequent. Most pupils seem to appreciate the opportunity to take a test on a small area of content immediately after it has been taught rather than at a later stage in the year' (p 134).

'When pupils were asked about the pressure that GASP brings the responses closely mirrored those on the subject of testing. 63% didn't feel that the scheme brings too much pressure whilst 16% agreed that it does. Again, when pupils were asked whether passing levels was too difficult, similar results were obtained. Commonly the same pupils agreed with both these criticisms so it may be that one source of the pressure is the feeling that levels are too difficult to achieve. Pupils at the lowest levels of achievement (below level 1 for first years and at level 1 or 2 for second years) are most likely to feel that levels are too difficult to achieve and less sure about the pressure that GASP brings' (p 135).

A sense of this pressure is evident in this comment from a second year pupil's essay: 'I don't like having to get things for levels because it puts you behind if you're not as good as everybody else. I've only passed my level one and everybody else is on three, or four or five' (p 135). There is a significant difference between the levels of satisfaction expressed by those students on the higher levels and those on lower levels. The majority of students interviewed felt that the award of certificates was an important aspect of the scheme and all said that they had kept their certificates.

One of the aspects that GASP students seem particularly enthusiastic about is the feedback they receive about progress. Just over a quarter of the essays referred in some way to the value of having short-term goals to work towards. The interviewed pupils tended to prefer practical explorations to non-practical and (ones) and in essays and interviews often referred to one or two practical explorations they had particularly enjoyed. Critical comments in essays outnumbered favourable comments almost 2 to 1, most pupils interviewed had reservations about some aspect of explorations. A common source of criticism was the amount of writing they involve or too difficult (p 136).

Most pupils do not feel confident that they understand the GASP system well. Second years feel they understand it better that the first years do and pupils at higher levels better than those at lower levels, but the only group in which a majority agreed that they understood the scheme well was that of the highest level second years (p 137).

Author conclusions

These first and second year pupils clearly have very positive attitudes towards GASP. For the majority it appears to be a popular scheme which gives direction to their work in science, confers a sense of achievement, provided feedback about progress and is seen as relevant to their future goals. The findings of this study add weight to the general feeling among teachers that most pupils enjoy working with GASP and are benefiting from it. There are three points of caution which should be noted, however. These are:

1. The scheme is new and novel to the pupils - attitudes may change as they get older
2. concern for those pupils, typically at the lowest levels of achievement, who feel GASP brings too much pressure and that levels are too difficult to achieve. In this study 16% fall into this category, the affect on these pupils of the constant reminders of their failure to progress which the GASP scheme gives, needs to be considered.
3. ‘Finally, there seems to be a case for making the scheme easier for pupils to understand. Descriptions and examples of skills, for example may be preferable to code letters and labels’ (p 137).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Johnston *et al.* (1993)	Teachers/ Teaching	English (reading, writing, speaking, listening)	Internal (for grading, in-school records, reporting to parents)	Exploration of relationships	High

Aims

To explore ways teachers assess children’s literate learning and their own professional effectiveness in teaching children to become literate. Particular interest in how teachers’ assessment techniques and frameworks were influenced by their knowledge, values and teaching situations.

Study design

50 teachers from Grade 1 through to Grade 12 from five school districts were interviewed. Of these, 21 teachers were from districts with ‘low control’ (encouraged not to use basal reading scheme), 13 were from districts with ‘high control’ (strict enforcement of basal reading scheme) and 16 teachers were from three districts of ‘medium control’ (somewhere between the two extremes).

Data collection

Semi-structured, open-ended interviews were used as a source of data. Interviews ranged from 45 minutes to two hours. Audio tapes were transcribed and analysed along the lines proposed by Spradley (1979).

Data analysis

Analysis was carried out by researchers not involved in data collection. Each researcher read five transcripts from two districts. Each researcher coded independently key terms and themes in transcripts. Then the group met and looked for common themes and applied tentative framework. All transcripts were read completely at least twice. After discussion, decoding and multiple readings the following domains emerged:

- context: level of control; elementary versus secondary
- knowledge of children's literature
- techniques used to assess literacy
- values expressed for teaching literacy
- nature of student's literacy development
- language used to describe literacy learning
- limitations on knowledge relating to assessment of literary development.

Coded segments were then tabulated by grade, nature of response and by teacher.

Author findings

Teachers' knowledge of children's literature

Teachers who knew more about children's literature were more likely to describe their students in terms of the students' selection and reading of literature.

The context of assessment: descriptions of student's literacy development provided by elementary teachers were almost all of greater length than those provided in upper grade (length used by researchers as a proxy for detail of description). Descriptions also differed in the extent to which they used 'subjective' language as opposed to 'objective' language: for example, the more controlling the context, the more distancing the terms, such as 'skill' and 'mastery'. In low control context, teachers began with a personal introduction to the student as an individual, not with levels; researchers found their descriptions more personal and more complex.

Knowledge of context

Teachers' knowledge of children's literature tended to influence their descriptive assessments of their students; this happened far less in the high-control context. High-control context also limited the extent to which teachers' knowledge of literature was reflected in their descriptions of the students.

Instructional goals and assessment strategies

Teachers' methods of assessing children's literacy development reflected their goals and values but also reflected the constraints under which they worked. There was considerable difference in the ways in which teachers kept track of students' work - in low control:

- running records
- regular written narratives
- ring binder with recorded comments on each student
- records of books read
- dialogue journals
- personalised assessment, often with tests playing a supportive role

In a high-controls context, keeping track of students' development through observation was less common. Serious testing was emphasised in every grade. Indicators of literacy in low-control included wide knowledge of authors and illustrators, types of language, choice of literary genre, etc. The major goal was that students should love literature; this was only noted by half of the teachers as a goal in a high-control context. For teachers in high-control context there was conflict between their beliefs about literacy and their beliefs about how to keep track of literacy development.

Teachers' self-assessments

The less controlled the environment, the more teachers reported paying attention to students' conversations about books as indicators of their own effectiveness. Parental feedback was valued by teachers in low-control context. These teachers tended to view standardised testing as providing taxpayers with numbers.

Teachers from high-control context mentioned test and standards.

All teachers in a low-control context spoke of a personal sense of knowing if they were succeeding; some of high-control teachers also felt this.

Teachers in a low-control context were more likely to admit to limitations in their own professional knowledge than teachers in a high-control context; this was also true of some teachers in high-control situations.

Author conclusions

Researchers found that teachers' knowledge and values were seen to influence their assessment of children's literacy learning and of the effectiveness of their own teaching.

Most teachers reported their primary source of knowledge was observation of students' behaviour and student talk.

When tests were emphasised by district, teachers' descriptions emphasised tests and they turned to tests for feedback about students and themselves.

Researchers found differences in school districts beyond high and low-control (e.g. urban and rural and size of the school district). The high-control district was urban and 20 times the size of suburban district that exerted least control.

Technical and bureaucratic control of teachers' instructional practice, partly through the use of tests, produced unelaborated descriptions characteristic of teachers who refer many students to special education.

Teachers in these systems feel less personally effective and notice children's development less: 'with 45 minutes once a day I don't know how to do it [teach literature]. I don't know that my students at the end of the year know how to think any better than they did when they walked in... the rewards are just not visible' (p114).

They believe these findings need to be explored in order to make informed decisions about approaching assessment.

Findings can be interpreted within the framework of construct and a consequential validity, viewing teachers essentially as assessment instruments. The data reveal the integral connection between construct and consequential validity in teachers' assessments of their students. Teachers who have a diminished sense of self-efficacy tend to rely on positional authority rather than personal authority in their teaching.

Data showed that teachers in urban districts were under more restrictive bureaucratic and technical control than those in suburban districts.

Researchers conclude 'our data suggests that the role of teacher knowledge in assessment and the role of assessment in teachers' knowledge should be taken very seriously. Conditions that constrain teachers' ability to assess their students and their practice in productive ways will not serve students well' (p 115).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Koretz *et al.* (1994)	Curriculum Teachers/ teaching	English (reading, writing, speaking, listening) Mathematics	External for accountability (high stakes for school)	Evaluation: naturally occurring	Medium

Aims

How effective has the Vermont programme been in meeting its goals? What goals appear to be in conflict in programmes of this type? Why is there a need for 'caution and moderate expectations'?

Study design

A variety of methods was used to examine programme implementation and impact; findings reported primarily reflect interviews of principals and teachers in a stratified random sample of approximately 80 schools.

Data collection

In the Vermont system in 1991-92, the mathematics portfolio scoring was conducted by teachers other than the students' own in regional meetings; in 1992-93, the scoring was done in a single state-wide meeting. In writing, teachers scored their own students' portfolios but, in 1993, this was done by other teachers, centrally.

A sample of scored portfolios was re-scored by second raters. The 'uniform test' in writing was limited to a single prompt. The mathematics tests were multiple choice. No details were given of interview schedules or questionnaires.

Data analysis

Statistical analysis of score data; descriptive for interview data

Author findings

Teachers and principals characterised the programme as a worthwhile burden. The programme demanded a lot of time and resources and imposed considerable stress: teachers reported an average of 30 hours a month working on mathematics portfolios. Administrators had to commit large resources as well, such as using funds for substitute teachers. It was not just time demands; there were also difficulties in finding appropriate tasks.

Educators found the programme a powerful and positive influence on instruction. Mathematics teachers reported more time given to problem solving and communication. Many noted changes in instructional practices (e.g. increase in time students spent working in pairs or small groups). About 50% of the teachers said they were more positive but fewer (43% grade 4, 27% grade 8) reported improvement in students' attitudes.

Indicative of positiveness was that, by the end of the first year, principals in roughly half the sample reported they had expanded use of portfolios beyond the grades and subjects that were part of the assessment programme. Wide variation was found in teachers' implementation of the programme, not surprising considering the nature of the reform and degree of autonomy granted by the programme. Although not unexpected, researchers found that differences could threaten validity of comparisons based on portfolio scores.

It was found that the Vermont portfolio programme was largely unsuccessful at that stage in meeting its goal of providing high-quality data about student performance. The writing assessment was hobbled by unreliable scoring, and mathematics assessment had problems of validity in confronting performance assessments in general.

Author conclusions

Results strongly suggest that the Vermont programme was more effective in meeting the second of its primary goals: that is, inducing changes in instruction. Although a powerful tool for encouraging teachers to change practice, success in this regard was incomplete. Also, positive effects of the programme have come at a steep price in time, money and stress. Non-financial costs were large.

Researchers conclude that, although the Vermont programme has shown promising effects on instruction and modest improvement in measurement quality in mathematics, a basic lesson to be drawn is the need for modest expectations, patience and ongoing evaluation with innovative large-scale performance assessments as a tool of educational reform. It was found that the 'Vermont experience has begun to make concrete the conflicts between the basic goals of this and similar programs and illustrates the need to make difficult compromises between them. Only additional time and information from other programs will clarify the extent to which the goals of the current movement can be met and the costs that will be required for meeting them...the Vermont experience to date highlights the need for caution and moderate expectations' (p 15).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
McCallum *et al.* (1993)	Teachers/ teaching	English (reading, writing, speaking, listening) Mathematics Science	Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

It can be implicitly inferred that the study authors attempt to elicit and describe implicit models of assessment practice used by year 2 teachers.

Study design

This is a qualitative study. Initially, participating teachers were interviewed about teacher assessment, then four consecutive days were spent in case study schools focusing on methods of teacher assessment and data collecting; schools that were not case study schools were then visited and data were collected. Postal questionnaires were then sent to all teachers.

These data were then used to create the different models of teacher assessment practices which were then validated through the use of vignettes sent to all teachers.

Data collection

1. Visits to all 32 schools involved to interview head teachers and Year 2 teachers about TA
2. Four consecutive days in each of the six case-study schools, focusing on methods of TA, including administering the 'quote sort' activity (included periods of observation)
3. Visits to the 18 non-case study schools, where the Year 2 teachers had not changed from the previous year using the 'quote sort' activity to focus on teacher assessment
4. Postal questionnaires sent to all Year 2 teachers about 1992 experiences of TA (and SATs); vignettes of our TA models sent to all Year 2 teachers as validation of our observation.

Note: The first component was considered not to have worked in eliciting detailed and explicit accounts of how teachers made their assessment, which was the reason for developing the 'quote sort' and carrying out stages 2 and 3.

Data analysis

Models emerged from analysis of all the data gathered. Particular information is given for the analysis of the 'quote sort'. First, a simple count was made of the number of teachers agreeing with each statement. Then a matrix was drawn up of teachers who agreed with each other in particular quotes. The final quantitative approach was to produce clusters of teachers using the cluster analysis utility on the 'data desk' statistical package. The detailed interview material was analysed using a constant comparative methods to produce groups of teachers with similar approaches/profiles. These groupings were matched against both the clusters and classroom observations of these teachers. The first tentative models were refined several times by the project team.

Author findings

The authors present five models of teacher assessment:

1. Critical intuitives divided into (i) children's needs ideologists (ii) tried and tested methodologists
2. Evidence gatherers
3. Systematic planners divided into (i) systematic assessors (ii) systematic integrators

Critical intuitives

i. Children's needs' ideologists show a great deal of confidence and can articulate arguments about assessment that defend a child-centred view of curriculum, teaching and learning.

ii. Tried and tested methodologists feel secure in modes of teaching and assessing practised before the ERA but are less confident in articulating what these are or their actual basis or uses for teaching or assessment purposes. Children's needs ideologists bitterly resent change imposed from outside, wanting to protect 'the human face of teaching' and their personal investment in it and worry that shifting to a focus on assessment could cause damage to children.

Change can threaten to invalidate long years of experience and some long serving tried and tested methodologists express a strong reluctance to accept the National Assessment model.

Evidence gatherers

These teachers have a basic belief in the primacy of teaching, rather than of assessing. Their main method of assessment relies on collecting evidence, which they only later evaluate. They have gone some way towards adapting to the requirements of National Assessment and they could be considered rational adaptors, in the sense that they have adapted in such a way as not to change their teaching: collecting evidence does not interfere with teaching practice. Another characteristic of evidence gatherers is that there is an increased awareness of National Assessment procedures and, in some cases, a degree of excitement brought by new developments.

Systematic assessment is seen as a threat to relationships with children. The teachers in this group have a fear of National Assessment interfering with their relationships with children, if they were to go over the top in adopting more systematic assessment practices. Recording or note taking on the spot was seen sceptically and acceptable only providing you just jot it down and do not make a thing of it. While some teachers acknowledged that recording on the spot was more accurate than reflecting back, they did not see it as a priority and had not incorporated it into their own assessment practice.

Systematic planners

Planning time specifically for assessment has become part of their practice and the planned assessment of groups and individuals informs future task design and classwork. This group of teachers plan for assessment on a systematic basis. This means that the teacher consciously devotes some part of the school week to assessing, and explicitly links the results of assessment to curriculum planning. Systematic planners see systematic diagnostic assessment as adding to their professionalism. They do not reject the notion that the child is at the heart of the learning process, nor that the 'whole child' is important. The balance of incentives seems for this group to outweigh the disincentives and the innovation of the National Assessment has offered them a sense of mastery, excitement and accomplishment. The systems they have devised can be seen as a stage in taking ownership of the innovation, enabling the teachers to retain a feeling of control.

Author conclusions

With no offered model of TA, it is perhaps not surprising that teachers came up with a range of approaches. These approaches were related to their espoused views of teaching and learning, their general style of organisation and their reaction to the imposition of the National Curriculum and its assessment. They were thus developing assessment practice in line with their general practice and philosophy of primary education. That it should have happened in this way is not surprising. What is particularly interesting is the relation between teaching, learning and assessment in the teachers' practice; the link between assessment and learning is a crucial one but is not generally widely addressed.

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Morgan (1996)	On teachers/ teaching	Mathematics	External for certification (high stakes for student)	Description	Low

Aims

‘To explore the ways in which secondary mathematics teachers read and assessed the reports of investigative work written by students as coursework at GCSE’

‘To provide knowledge to help both teachers and students to produce coursework texts that are more likely to be evaluated positively’

‘To raise questions about the validity of the assessment itself....and address the relationship between the examination process and the curriculum reforms that it was intended to support and encourage’ (p 356)

Study design

Eleven secondary mathematics teachers were asked to talk aloud as they read and assessed examples of coursework. By implication, this commentary was recorded and as least partly transcribed. Narrative accounts were given of approaches to various part of the task by two teachers in particular, with reference to others at times.

Data collection

‘During individual task-based interviews, each teacher was asked to read and assess a set of three students’ coursework texts reporting work on the same investigative task, talking aloud as they did so, and eventually to rank the set of texts. The student texts had been selected to display a variety of characteristics related to the forms of communication used; they also varied to some extent in their approaches and solutions to the problems tackled, although all had made approximately comparable progress through the investigation and had been given very similar marks by their own teachers’ (p 356).

Data analysis

‘The analysis of the data was achieved through close attention to the text of the transcripts, making use of linguistic cues or ‘danger signals’ (Jensen, 1989) to identify issues of significance to the participants and considering the narratives produced as windows into the interpretative resources used by teachers to make sense of their own practice and to justify their judgements’ (p 357).

Author findings

Quotations from two teachers show that one (Joan) tends to ‘read in’ to what the student has written in order to understand it, whilst the other (Fiona) is focusing on what it not there.

‘Joan’s initial position while reading this section of the text is an interested reader, seeking to form an understanding of both the meaning and validity of the text. It is only once she has made sense of and validated the mathematics that she takes on the assessor role, evaluating the form and identifying mistakes.

‘Fiona, by contrast, does not attempt to make mathematical sense of Richard’s formula until after she has considered and evaluation other features of his work’ (p 360).

The teachers’ differing stances are seen as, on the one hand, taking a teacher’s role and on the other an assessor’s role. ‘There is a tension between the rigour of the examiner, for whom the assessment criteria determines unproblematically a decision about the value of a piece of work, and the wish of a teacher, acting as advocate on behalf of her pupil, feels, as Joan did, that the missing evidence might have been available in the classroom. The problems are Richard’s but they are also Fiona’s problems in resolving her two roles’ (p 361). The tension between taking on an examiner role and acting as teacher/advocate is a familiar one for teachers involved in any summative assessment. It has been observed that, while working with students engaged in coursework tasks, teachers struggle to be ‘fair’ both to their own students and to the integrity of the assessment system (p 362).

There is a contrast between the assessor as technician (who only needs to be able to apply the criteria and may not understand them?) and the assessor as interpreter whose content knowledge is key to understanding what students are trying to do. The differences identified between Fiona and Joan reflected varying positions taken by the other teachers interviewed. 'In particular, each teacher's practice could be described predominantly either as comparing each student's text with an imaginary 'ideal' or as building up an imaginary picture of the student and his or her activity' (p 367).

'The general strategies used by the teachers interviewed may be crudely classified as focusing on the student or focusing on the text' (p368).

'In spite of her negative evaluation of many features of Steven's text and her identification of weaknesses, Jenny had gained an impression of his 'ability' that caused her to state that he should achieve a C grade....The quality of C-ish was attributed to the student rather than to the particular piece of work' (p369).

It appears from these findings that even when teachers only have the written outcome to go by, they may still allow impressions of 'general ability' to influence their assessment.

Impacts

On assessments

In most cases, judgements based on a construction of a students' personal characteristics and those based on a search for the fulfilment of criteria or a comparison with a 'ideal' text gave rise to similar rankings of the student texts.

On teaching and students

The most difficult texts to assess were those differing from the stereotypical. This leads the author to suggest that 'the examination process itself thus discourages creative or unusual ways of thinking. Exciting, innovative, creative mathematician-students may well work in non-routine ways and hence need to develop non-routine ways of communicating their work. While these characteristics may lead some teachers to value the work, they are also likely to give rise to conflicts both between different teachers and between the different positions potentially adopted by a single teacher. Such conflicts jeopardise the assessment system's requirements for reliability and simplicity. Teachers must, therefore, be under pressure to avoid them while preparing their students for the coursework examination' (p 372).

But similar rankings of students' texts were noted, suggesting that the different approaches to assessment that emerged do not have major effects on the outcome of the examination process.

Author conclusions

The ideals of creativity, student ownership of their work and participation in genuine mathematical activity that gave rise to the development of investigative work in maths and its subsequent institutionalisation as part of GCSE coursework was (thus) put at risk by the very assessment system that was intended to encourage them.

'In summative assessment like the GCSE, in which the main purpose is to sort or rank students, it may be acceptable that different teachers arrive at the same or essentially similar conclusions by different means. This, however, is of very dubious value in cases where the assessment is also intended to be used formatively' (p 372).

The author suggests that current training for coursework assessment emphasises reliability and agreement with other teachers and 'further reinforced the trend towards standardisation in the tasks set and in the students' work. It is possible that in-service education, taking examples of assessment practices like those presented in this paper as a staring point for critical reflection on the assessment process, might have more positive effects' (p 372).

'Although it is true that, following the introduction of GCSE, most UK maths teachers now use 'investigations' in their classrooms, these often bear little resemblance to the original ideals of investigative work, being little more than routine exercises…. There is, therefore, a need for further critical examination of the principle of assessment-led curriculum development and of the means of implementing it in particular cases' (p 373).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Pilcher (1994)	Students	English (reading, writing, speaking, listening) Mathematics	Internal (for grading, in-school records, reporting to parents)	Description	High

Aims

'This study investigated what students and parents perceive grades to represent' (p 73). The hypothesis was that 'the meaning these three audiences associate with grades depends on how grades are interpreted, used, and valued as well as the social consequences students face when grades are used to make decisions abut them' (p 73).

Study design

Six case studies were conducted, using data gathered by interview and from documentation. Individual interviews were carried out. Documentation was used for 'verifying and guiding responses'. Results were analysed to enable comparisons and contrasts to be made within cases (each comprising the English and mathematics teachers, the student and parent) and between cases. Interviews were transcribed and used to develop narratives for each participant on how they interpreted and used the grades assigned to the student. From these narratives, the information was organised into broad categories. The narratives were also used to test the hypothesised 'grading equation for teachers and to develop the grading equations for students and teachers' (p 71).

Data collection

One-to one interviews, presumably with examples of grades to discuss. No information about duration or location of interviews. No information about how documentation was located or selected.

Data analysis

Qualitative analysis of the data carried out as follows:

'I analysed within-case and between-case comparisons and contrasts. Comparisons and contrasts of responses were made within each student, parent, and teacher combination and among all students, all parents and all teachers' (p74).

'The data analysis included four phases. During the first phase, interviews with all participants were recorded on take and transcribed...In the second phase I wrote narratives for the 24 participants...In the third phase, information from the narratives was organised into broad categories on coloured index cards by using analytic techniques pattern matching and analytic induction. The categories were are follows: what grades mean for the case study student, what grades assigned to students in the same class as the case students represent, and what grades assigned to students in other classes represent. In the fourth phase, I placed the index cards into six categories that represented each case to view patterns of responses between all student, teacher and parent combinations. To determine the meaning of grades, the hypothesised grading equation for teachers was tested and the grading equations for students and teachers were developed' (p 74-75).

Author findings

The main findings were focused on the six week grades, which incorporated test scores, assignments during that period.

'English teachers claimed that they grade writing assignments based on a student's perceived writing ability...even though the math teachers of the two above-average students perceived they graded students on solely achievement factors, both assigned a zero to incomplete homework and distributed extra credit points for optional graded assignments' (p 76).

Narratives for each student indicated the following: 'teachers were more negative towards students who did not complete assignments or scored low on tests (than those who were considered 'bright'). Teachers perceived these students deserved a low grade' (p 77).

'…to Alice, a grade represented how much a teacher liked a student and the amount of work a student completed. ...She received higher grades in the classes in which she perceived the teachers liked her, which usually meant providing her with special attention...Alice's teachers realised that when Alice had a reason for making good grades, she applied the effort to do so....Alice's father used her grades to "judge whether or not Alice needed help in a subject". Alice admitted that most of the time she did not need help. She only needed to complete the assignments' (pp 79-80).

'This study suggest that grades represent a combination of achievement of course content, ability level of students and effort applied in class. A high grade means some combination of high test scores, high grades on writing assignments, and applied effort. A failing grade means total lack of effort. Even students who fail tests can usually receive a passing grade if they apply effort' (p 80).

'In essence, some students are penalised by grades whereas others are rewarded...' (p80).

'Students seemed to be more negative about maths teachers grading practices than those of English teachers. Students were more accepting of point deductions for late assignments in English than in maths...Maths teachers' grading procedures were perceived to be more concrete; that is, they judged student work as right or wrong. English teachers adjusted achievement grades by considering a student's writing ability' (p 81).

'There were also more maths assignments than English ones, so 'English teachers' grading did not require the daily motivation of students to turn in assignments' (p 81).

Different grading equations were identified from the study (see p 81-2).

'Although both math and English teachers did not assign a score to attitude, 'teachers' comments about adjustments of grades for particular students indicated that teachers made inferences about attitudes when assigning grades' (p 82).

Students were aware of how teachers were grading them and so interpreted their performance in similar ways to the way the grades were assigned.

'Parents did not interpret grades as teachers assigned them. Parents perceived grades reflected their child's achievement level' (p 83).

Students had varying views on how grades should be assigned: three wanted teachers to include effort (but only to increase scores, not to lower them).

It was clear that grades were used by teachers and students as extrinsic motivation. Teachers 'preferred that students apply effort by their own will. However, they admitted that they had to implement grading procedures that would reward student effort and punish students for lack of effort, otherwise students would not apply themselves...Grades, therefore have more extrinsic rather than intrinsic value to students. Teachers and parents used grades as extrinsic motivators to encourage students to perform' (p 85).

Author conclusions

'Teachers value the extent to which students do what they are instructed to do. Parents value the degree to which their child has mastered the content as compared to other children. Teachers and parents use grades to extrinsically motivate students to complete assignments and perform in class. A grade, therefore, is a token distributed to students for their performance in class. Students who have a good attitude, are responsible and have high quantitative scores receive a high reward' (p85).

A problem with using grades to control students occurs when students resist doing something...or are not responding to punishment by their parents. (Extrinsic motivation only works if the student places a high value on grades.)

'This study indicates that teachers and parents use rewards and coercive power to influence the grades students receive....French and Raven (1960) claimed that both reward and coercive power are the least effective types of power in producing positive student behaviour towards learning' (p 86).

'This study suggests that in their current and recommended states, grades are more harmful than beneficial to student learning. Using grades to control student behaviours does not teach students to value learning' (p 87).

The author suggests that educational reforms in assessment need to focus on practical methods for reporting as well as on alternative methods of assessment: '...the purpose of grades needs to be questioned to develop grading procedures that do not interfere with a student's intrinsic value of learning. The positive and negative consequences students face from the value others attach to grades must be addressed' (p 87).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Stables (1992)	Students Teachers/teaching	English (reading, writing, speaking, listening)	National Curriculum Teacher Assessment of oracy	Evaluation: naturally occurring	Low

Aims

To investigate teacher assessment in English, mathematics and science at Key Stage 3 of the National Curriculum. This study reports only on that part relating to the assessment of speaking and listening at KS 3.

Study design

The study was a cross-sectional study of how teachers in four classes went about assessing speaking and listening. Data were collected by observation of lessons and interviews with teachers and discussion at periodic network meetings. Most interpretation is based on classroom observations; no other data are reported systematically.

Data collection

Observation of four lessons. Number of one-to-one interviews not specified. Data collected from two network meetings. No details of observation procedures given. Results suggested that these were unstructured narrative descriptions.

Data analysis

Not stated

Author findings

Listening was found to present particular problems, especially when work is undertaken in pairs or groups. The data revealed a number of areas in which current assessment methods were seen to lacking clarity.

The issues arising were reported in six areas:

1. There was a tendency to see English as a 'process' rather than a 'product' subject. In situations involving groups and pairs, it was not at all easy to assess an individual student's contribution to the process by observing the product (p 111).

   The impact reported was that teachers were more aware of the need to give attention to the setting of the task in order to allow each student to perform well (see p 111). To assess what was required demands an increased awareness on the part of both of pupil and teacher of what those competencies are and how they can be developed. Sadly, in the case of spoken English, they may still be a great deal of work to be done to develop teachers' understanding of the processes themselves (p 112)
2. More attention needs to be given to the setting of the task in order to allow each student to perform well (p 111).
3. There was an expressed concern that the National Curriculum never called for group responsibility. There was a need for training in the development of group discussion and some called for rotation of groups among students (p 112).

4. It is very difficult for the teacher to be an assessor while actually teaching (p 113).
5. There was concern that assessment demands might lead the teaching rather than the other way round. The network members held firmly to a belief in the open-ended, differentiation-by-outcome nature of good English teaching (p 113).
6. Effective teacher assessment will partly mean having the evidence of pupils' progress in a sufficiently clear and detailed form to be able to challenge Standard Assessment Task (SAT) results at moderation.

Author conclusions

Teachers have been left in the difficult position of having to keep records with very few indicators of the kinds of evidence they will have to produce. It would seem clear, however, that the more concrete and specific the student records are in terms of the assessment demands of the National Curriculum (i.e. the Levels and Statements of Attainment), and in terms of the recording or citation of students' speech, the more incontrovertible they will seem as evidence of students' progress at moderation (p 115).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Valencia and Au (1997)	Curriculum Teachers/ teaching	English (reading, writing, speaking, listening)	Formative and summative Internal (for grading, in-school records, reporting to parents)	Evaluation: naturally occurring	Medium

Aims

'Specifically, we explore (a) the potential of literacy portfolios from different sites to capture curriculum outcomes that are both authentic and aligned with instruction, (b) teachers' ability to interpret and evaluate portfolio evidence from more than one site, and (c) what teachers learn about literacy instruction and assessment through the process of cross-site collaboration' (p 3).

Study design

Teachers at two sites, where different portfolio assessments were in operation, were interviewed before a series of cross-site meetings. In these 2–3 day meetings, they observed in each others' classrooms and then in groups reviewed and evaluated portfolios from each site, using first the criteria developed and used at one and then the other site. Finally they developed common criteria and used these to evaluate the portfolios. After these meetings they were interviewed again.

Data collection

Portfolios were evaluated at each meeting and were analysed to identify variability across sites. Pieces were coded according to the major disciplinary focus, the type of artefact (student work, anecdotal notes and checklists, students' self-selected pieces, goals or self-evaluation, etc.). Ratings by the teachers using the criteria for each site and common criteria developed by the teachers were collected.

Audio-tapes of the meeting transactions and interview data from pre- and post- meeting interviews were also used.

Data analysis

Inter-rater agreements for the teachers' rating of the portfolios were calculated (p 15). Contents of the portfolios were categorised and codes according to: (i) major disciplinary focus (reading, writing and 'other subjects'); (ii) type of artefact (student work; anecdotal notes and checklists, other, student selected pieces, goals or self-evaluation); and (iii) parent input (questionnaires, comment).

No information about the analysis of interview data or how the recordings of meetings were used.

Author findings

Portfolio content analysis

There were similarities in terms of average number of pieces in each portfolio. There were consistent differences across sites in the focus of the pieces of work. However, there was more consistency in both discipline and type of artefact among portfolios from the same classroom than among portfolios from the same site.

‘While reviewing each other’s portfolios at the first cross-site meeting, teachers spontaneously discussed types of evidence that were new or interesting to them. ... A review of portfolio contents at the second cross-site meeting revealed that teachers from both sites had incorporated some of the teachings strategies and portfolio artefacts they had learned from each other’ (p 18).

‘We should note, however, that the teachers were as struck by the similarities in their philosophies and instructional emphasis as by the differences in the portfolio contents’ (p 18).

Overall, the content analyses suggest that portfolios contained high-quality authentic samples and records of students’ reading and writing. They reflected the underlying outcomes and portfolio models of each site as well as the emphasis of individual teachers. However, evidence of some outcomes and some types of work, particularly reading outcomes and discussions, was apparently difficult to document using a portfolio, resulting in limited information about particular outcomes. Nevertheless, the process of reviewing and discussing portfolio contents provided teachers with ideas about both assessment and instruction that carried over into their own classrooms and their students’ portfolios.

Where the evaluation criteria demanded very specific evidence, there was more missing evidence.

Evaluation process and results

The relatively strong inter-rater agreement suggests promising possibilities for reporting results for groups of students across sites. Such agreement is remarkable because the portfolios did not contain similar pre-specified pieces; they were developed to be most useful at the local district and individual classroom level. As a result, there was substantial variation in the contents across portfolios from different sites. In addition, because most of the pieces in the portfolio were complex, authentic examples of reading and writing, individual pieces often contained evidence of several outcomes or benchmarks.

Particular attention was paid to ‘missing data’ (criteria for which there was no evidence). The findings indicate that, even when teachers decide the evaluation criteria locally, ‘if teachers or students are unfamiliar with the evaluation criteria ahead of time, them may not select work required by the criteria’.

‘Our analysis suggests that teachers did not have a bias when interpreting and rating portfolios. Bellevue teachers and Kamehameha Elementary Education Program (KEEP) teachers did not consistently rate their own portfolios higher or lower that the teachers from the other site...In light of concern about fairness of new assessment, this finding is especially promising. Not only were the students from very different cultural backgrounds, but the teachers (raters) were as well’ (p 23).

Several factors contributed to the consistency in rating among teachers:

Teachers shared common conceptual understandings about literacy learning... these understandings were further enhanced through the classroom visits, the across-site portfolio discussions and the evaluation process that were part of this study. Integral to this professional collegiality was the low stakes nature of the project.

Impact on teachers and teaching

The study takes two teachers as examples of the impact (professional development) of the project processes.

‘At the conclusion of the project, Sue noted that she had started being more specific with her students about the kinds of evidence needed to document their involvement with the writing process....She had her students evaluate their own writing using an evaluation rubric as a means of teaching them about standards for literacy performance. She had also become more aware in her own classroom of the quality of students’ responses during literature discussions’ (p26).

‘A portfolio content analysis served to verify that Sue had succeeded in making several changes’ (p 26).

Of the second teacher: ‘Nora (a KEEP teacher) indicated that involvement in portfolio evaluation and development of the Common evaluation rubric had broadened her perspective on portfolio contents and organization. She gained confidence in her ability to understand descriptors and evaluate individual students’ portfolios...Nora observed that she had become aware of the importance of students’ self-evaluation and reflection and regular portfolio visits, central features of the Bellevue portfolio system’ (p 27).

‘As these case examples indicate, the teachers in this project gained specific ideas for improving portfolio assessment in their classrooms. Their study, rating, and discussion of each other’s portfolios heightened their concerns about documenting students’ literacy learning in as complete and sensitive a matter as possible. Teachers seemed to be adding to their notions of what should be included in a portfolio and broadening their thinking about important dimensions of students performance’ (p 28).

Author conclusions

1. Portfolio evidence, the evaluation process and professional development must all three be in place if portfolios are to improve instruction, student achievement and classroom-based assessment (p 29).
2. As teachers closely examine students’ work from their own and other classrooms they clarify important learning outcomes and learn to interpret student performance based on multiple forms of evidence (p 29).
3. Using portfolios for evaluation forces teachers to rely on evidence for their decisions; it grounds evaluation in concrete data (p 30).
4. Had the requirements for portfolio implementation been more prescriptive, they would have been too removed to be meaningful and useful to teachers.

‘Our data and experiences suggest that, as teachers closely examine students’ work from their own and other classrooms, they clarify important learning outcomes and learn to interpret student performance based on multiple forms of evidence. Working collaboratively and discussing portfolio artefacts encourage teachers to re-examine their knowledge, assumptions and misconceptions about teaching and assessment practices. Conversely, teachers’ understandings of curriculum and instruction are reflected in portfolio evidence. We can see their strengths and weaknesses, their priorities, and the opportunities their provide for their students. As teachers grow and change through professional development, their portfolio evidence begins to change as well. This slow, iterative process is likely to produce meaningful sustainable changes in both portfolio evidence and in underlying classroom instruction’ (p 30).

‘Supportive internal conditions are essential to sustaining an effective portfolio system. In this study, the local conditions within each site supported the emerging portfolio systems. Stakes were low for teachers; district interest, support and commitment were high’ (p 31).

Without supportive internal and external conditions, the results of this study certainly would have been different.… However, with all elements in place, we believe the process is a strong one - it has system validity (Frederiksen and Collins, 1989; Messick, 1994). This assessment process promotes value changes in teaching and learning; this is the ultimate goals of assessment’ (p 32).

Study	Object of impact reported	Achievement assessed	Use of result	Study type	Overall weight of evidence
Whetton *et al.* (1991)	Teachers/ teaching	English, Mathematics Science Other: Welsh as a first language	Internal (for grading, in-school records, reporting to parents)	Description	Medium

Aims

The purpose of this report is to provide evidence about the methods teachers have used to make their assessments for the first full run of National Curriculum Assessment (NCA) for Key Stage 1 and the support they received.

Study design

The appendix makes clear that evidence on which the report is based comes from two sources: a questionnaire study and a series of case studies. Both these are part of an ongoing evaluation of the SAT and the operation of NCA as a whole. Questionnaires were short and were designed to elicit differential information about attainment targets as a whole. Schools were selected from the NFER register of schools.

Data collection

Questionnaire survey and case study interviews/visits

Data analysis

Descriptive statistics to give percentages of answers of different kinds

Author findings

1. Support for teachers

1.1 Central support

Provided through the School Examination and Assessment Council (SEAC) – distance-learning material was sent direct to schools: ‘School Assessment Folder’ (SAF) in November 1990 (included ‘Children’s Work Assessed’ (CWA)) and ‘A Guide to Teacher Assessment’ (GTA) sent in January 1990. Also SAT given to schools in March 91, had influence on TA.

Many found used SEAC materials to assist with TA. SAF was valuable for most. CWA was useful, although some found too much emphasis on Level 3. Fewer teachers found GTA useful.

It was found that the majority of schools had not used materials as they found the content complex and inaccessible.

1.2 LEA training days (INSET)

Quality of INSET was extremely variable. This was most successful where trainers were knowledgeable, confident and positive.

Some teachers had not attended INSET relating to TA. Many found the days to be rushed.

Most value from days focusing on specifics; incorporating time for small group discussions; containing opportunity for Y2 teachers to spend time on their own concerns.

Least value from days targeted at more than one audience; delivered by ill-briefed personnel; badly planned; lacking in recognition of previous INSET; lacking in substance.

The opportunity to meet other teachers in small groups was very valuable.

Case study evidence suggests that quality of INSET support depended less on whether LEA was involved in the pilot, than on the structural arrangements for INSET within the LEA, particularly cluster groups.

1.3 Support within schools

Y2 felt supported in schools with carefully drawn up policies and recording systems which had been in place for some time.

Many teachers provided with 'non-contact' time as support. Some LEAs funded supply cover for this.

1.4 Experience of the pilot

Case studies showed that experience of the 1990 pilot was by far the most effective form of support for Y2 teachers.

2. Recording evidence of attainment

2.1 Using statements of attainment

The majority of teachers attempted to assess and record against Statement of Attainment (SoA) (96%).

Different schools had different recording systems. In general, teachers were struggling with the issue of how many times to assess each SoA, and some teachers put themselves under lots of pressure due to too much assessing.

2.2 Types of evidence

Questionnaires asked teachers whether they used teacher record, students' work, both or none. Substantial differences were found (recorded in Figure 2.1, p 19) but majority used both for their TA.

Study found one in eight teachers made judgements using no recorded evidence.

2.3 Use of non-statutory materials

One-third used Speaking and Listening materials provided, and 94% found it useful. Reasons for not using it included lack of time, not needed as records complete enough, considered to be of poor quality, difficult to manage, and that it arrived too late to use.

3. Using Document A

Questionnaire only asked about use, not ease of use. Only 1% (three schools) used D code (disapplied). 19% used N code (non-assessed), most often for children coming from another school without records.

4. Difficulties experienced by teachers

4.1 Non-coverage of programmes of study

Case studies indicated that most experienced problems completing Document A as they had not covered all aspects of the programme by 31 March.

Coverage of attainment targets related to teachers' ratings of the difficulty of making judgements about them.

Most teachers struggled with the problem of non-coverage of programmes of study by attempting to teach and assess during Spring term all targets not previously covered. Result was increased workload, distortion of curriculum and anxiety.

4.2 Programmes of study covered in Year 1

Case studies confirmed that difficulties were most common for attainment targets of science profile component 2. Problems were magnified in the case of SoA with multiple attributes. Improved record-keeping systems will go some way to alleviating problems reported above. Pilot schools in general had well-developed record-keeping systems and experienced fewer difficulties.

4.3 AT and SoA

Teachers were asked to rate how easy it was to make judgements about each of the attainment targets. Some process-based ATs (Ma1, Ma9, Sc1) together with some specific knowledge-based ATs were judged to be the most difficult. EN5 (handwriting) was considered the easiest.

4.4 Students with SEN, not fully fluent or from different ethnic backgrounds

Very few teachers expressed concerns of assessment of special groups of students, although many were concerned about reporting to parents that children had not yet reached level 1.

4.5 Retaining evidence.

The majority of case study teachers were uncertain and confused about retaining tangible evidence of students' attainment.

It was evident that teachers need to be given consistent advice on extent to which they need to retain evidence.

4.6 Unfamiliarity, workload and teacher attitudes

Many teachers commented upon the excessive workload they had to undertake to complete TA. Case studies showed that much of workload was caused by inefficient systems for collecting and recording evidence of attainment. Unfamiliarity with NCA also contributed.

Those schools which had been involved in 1990 pilot approached TA with greater calm and confidence. Their assessment and recording policies tended to work more efficiently, resulting in less stress and a more manageable workload.

Teachers expressed greatest difficulty with the process-based attainment targets in maths and science, together with specific knowledge-based attainment targets in mathematics and science profile components 2. In English, speaking, listening and writing were most difficult; handwriting was easiest.

In interviews, teachers expressed difficulties because of vagueness and generality in wording of SoA. They stated that accumulated banks of SATs would be of great assistance in future communicating of 'what is required'.

Teachers varied in degree of confidence ascribed to TA results. In several cases, confidence had increased as a result of agreement trialling experience.

In general teachers had more confidence in their TA for those aspects of children's attainment which are manifested within context of classroom work.

Author conclusions

No conclusions given separately from the results

Study	**Object of impact reported**	**Achievement assessed**	**Use of result**	**Study type**	**Overall weight of evidence**
Yung B (2002)	Students Teachers/ teaching	Science	External for certification (high stakes for student)	Description	High

Aims

'The prime aim of this study was to find out, from the perspectives of the actors [the three teachers] what was happening in their classrooms and why they acted in certain ways' [when the Teacher Assessment Scheme was being used to assess students' practical work] (p 101).

Study design

A study of three cases of teachers assessing students' practical work in biology using the TAS, with data collected by interview and classroom observation and analysed qualitatively.

Data collection

Lessons were recorded using a microphone on the teacher. Lesson observation of at least four practical sessions of 2–4 hours; other lessons related to practical observed (number not specified). Observations were supplemented by post-lesson interviews that probed the teacher's pedagogical decisions as well as their associated thinking and beliefs.

Data analysis

Interpretations of the data were constructed using a method similar to that described by Erickson (1986 p 119) focusing on the immediate and local meanings of actions, as defined from the actors' point of view.

This was done by reading the transcripts in relation to the classroom data and linking them to what the teachers saw as significant in what they did in the classrooms' (p 101).

Author findings

The results are presented as case studies for each teacher, indicating the practice in using the TAS, the teacher's explanation of it and the interpretation given by the author.

Case 1 was a teacher who kept very closely to the TAS regulations. He followed the suggestions for using the assessment formatively both in his teaching and in discussions later with students. He explained this in terms of making him a better teacher and the interpretation was that he did this in his own interests rather than the students'. His approach to the regulations was 'readerly'. 'To conclude, for Ivor, despite his stated desire to improve both his teaching and his students' learning...the introduction of the TAS with its many regulations was seen as imposing severe constraints upon his professional autonomy. Under such circumstances teacher professionalism was severely compromised as Ivor struggled to make sense of his changing role and responsibilities both as an assessor and as a teacher. Most importantly, underpinning his 'readerly' interpretation of the TAS regulations was his consciousness of 'protecting himself' rather than defending his students' interests' (Yung, 2002, p 104).

Case 2 was a teacher who took a 'writerly' approach to the regulations. 'With a high level of professional confidence, Carl was able to impose his own professional interpretations on the TAS regulations, to balance its demands against other professional priorities and to exploit to the maximum what remained of his professional autonomy' (p 105). Thus he allowed students to work in ways that were more natural (in groups, conferring with each other and looking up references): 'the most significant influence on Carl's teaching was not the formal apparatus of external obligation and controls imposed on him by the TAS regulations, but his personal sense of professional obligation to offer students an all-round education' (p 107).

Case study 3 was a teacher who also did not keep exactly to the regulations, but with a different motivation: of making sure that students got high grades. He was exam-oriented in focusing just on what would help students to get high marks. 'Evidence suggests that Eddy's consciousness was directed towards his own self-interest rather than his students' learning. This in turn led to his own low level of professional confidence and, hence to a seemingly writerly but actually 'readerly' interpretation of the regulations' (p 111).

Teachers who exhibit low levels of professional confidence can feel insecure about educational change; professional consciousness can be directed by self interest and self protection rather than the good of, and defence of, the students; feeling of powerlessness and resignation that a scheme that may be misguided was imposed on them leads to severe constraints; struggle to make sense of role and responsibility as teacher and assessor; lower levels of professional confidence may lead them to indulge in unprofessional practices; fear of getting it wrong in front of superiors and examiners; wish to enhance own image to others rather than beneficial effect for students; reactive rather than proactive.

Teachers who exhibit high professional confidence see change as an opportunity to reassess their own pedagogic methods in order to enhance student learning and welfare; they have strong beliefs in student participation and have a personal sense of professional obligation to provide students with all round education; they also have a strong commitment to own educational goals, confidence in their own ability to be proactive rather than reactive to change, and use their own professional interpretations on regulations.

Therefore, professional consciousness and teachers' responses to TAS was a matter of each teacher's personal choice, being guided by their own beliefs and personal theories about teaching and learning.

A larger study found that, whilst all teachers were constrained to a greater/lesser degree by TAS regulations, some were more ready to follow external prescriptions even when judged to be misguided. However, where teachers have a high sense of professional confidence and higher professional consciousness which is guided by professional confidence, they seek to integrate TAS requirements within their own professional autonomy to work for the best interests of the student by adopting a critical stance to policy change.

A teacher's professional consciousness has a strong bearing on what stance he/she would adopt towards policy change, whether control of own teaching can be exercised, and whether it can be made to work in the interests of students.

Author conclusions

‘Teachers who adopt a critical stance to policy change are able to exercise control over their own teaching…that the variations in how teachers adopt such a critical stance is a function of the professional confidence, which in turn is guided by their professions consciousness. The current study extends this relationship by pointing to the importance of teachers having their professional consciousness directed towards protecting students’ interests rather than their own...’ (p 115).

The author indicates ‘serious implications for teacher education and teacher pd [professional development] in general, and in the area of preparing teachers for more school-based assessment in particular’: that is, relevance for the conditions that enable teacher assessment to be implemented in the best interests of teaching and learning.

Overall, study data are consistent with the literature that teachers who adopt a critical stance to policy change are able to exercise control on own teaching. Variations in how teachers adopt this position is a function of own professional confidence which is guided by professional consciousness.