PROTOCOL

July 2004

Evidence for Policy and Practice Information and Co-ordinating Centre

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

Protocol written 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

REVIEW GROUP DETAILS

Institutional base

Graduate School of Education
University of Bristol
35, Berkeley Square
Bristol BS8 1JA

Tel: 0117 928 7129

Contact address for co-ordinator

Haymount Coach House
Bridgend
Duns
Berwickshire TD11 3DJ

E-mail: wynne@torphin.freeserve.co.uk
Tel: +44(0)20 7612 6131 / 01361 884710
Fax: +44(0)1361 884013

Advisory structure for the review

Members of the Assessment Review Group (ARG) and their affiliations

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
Professor Judy Sebba, University of Sussex
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 the Association of Assessment Inspectors and Advisers (AAIA)
Mr R Bevan, Deputy Head Teacher, King Edward VI Grammar School, Chelmsford
Ms P Rayner, Link Inspector for Primary Education, Nottinghamshire

International experts (advisers to the Assessment and Learning Research Synthesis Group, ALRSG)
Dr Steven Bakker, Educational Testing Service (ETS) International, The Netherlands
Dr Dennis Bartels, Director, President, TERC, Cambridge, Mass., USA
Professor Lorrie Shepard, President, American Educational Research Association (AERA), 1999-2000, University of Colorado
Professor Eva Baker, Co-Director of the Center for Research on Evaluation, Standards and Student Testing (CRESST), University of California, USA

Dr T Crooks, Director, Educational Assessment Research Unit (EARU),
University of Otago, Dunedin, New Zealand
Professor Dylan Wiliam, Educational Testing Service

All the UK-based members of the Review Group act in an advisory role. Meetings of the whole group will be arranged to coincide with the main decision points:

- refining the protocol and confirming that the range of searches is as extensive as possible
- reviewing the map of the research and identifying exclusion and inclusion criteria for selecting studies for in-depth review
- discussing the synthesis of findings
- reviewing the draft report

Overseas advisers will be consulted in general by mail, although opportunities for face-to-face consultation will be taken where available through conference or other networks.

The main work of the review will be carried out by Professor Wynne Harlen, between February and September 2004, with publication of the report completed by January 2005. Two research assistants will be employed for the period May to July, when keywording and data-extraction will be taking place. Other members of the Review Group will take an active part in the review processes, according to their availability, as well as acting in an advisory role.

ACKNOWLEDGEMENTS AND CONFLICTS OF INTEREST

Funding for the review from the Department for Education and Skills (DfES)-funded Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) is gratefully acknowledged. This funding is supporting the conduct of the review in accordance with the procedures and using the tools being adapted and developed for synthesis of educational research and providing for:

- the review co-ordinator (Professor Wynne Harlen, 25% fte)
- research assistance, May to July
- supply cover for one teacher to attend ALRSG meetings
- travel costs for ALRSG members
- cost of stationary, postage, phone
- library charges

There are no conflicts of interest for ALSRG members.

TABLE OF CONTENTS

1. BACKGROUND

1.1 Previous work of the Assessment and Learning Research Synthesis Group (ALRSG)

The ALRSG was created as one of the first wave of the Evidence for Policy and Practice Information and Co-ordinating Centre (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 Information and Communication Technologies (ICT) for assessment of creative and critical thinking skills, and was 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 (Black and Wiliam, 1998; Crooks, 1988;) that, on the one hand, formative assessment can raise standards of achievement, and that, on the other, 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 had found that testing has a negative impact on motivation for learning. There were indications that policy-makers' attention was turning to considering greater use of assessment by teachers as an alternative to testing. However, there were concerns about the dependability and effects of assessment by teachers used for summative purposes. Thus the third and fourth reviews were set up to seek evidence in relation to these concerns. The third review, concerned with the reliability and validity of assessment by teachers for summative purposes, was published in March 2004. This 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.

1.2 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; and
(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, a teacher feels pressure of a different kind in being both teacher and assessor, a dual role which some feel 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 – and there is 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.

Counter opinions 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 the process of 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; National Council on Education Standards and Testing (NCEST), 1992). Providing that the process is an open one, where 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 (p 8).

It is the intention of this review, by studying the circumstances in which the advantages of using teachers' assessment for summative purposes can be achieved without too many of the disadvantages, to identify implications for policy and practice.

1.3 Definitional and conceptual issues

1.3.1 Educational assessment

Assessment in the context of education involves deciding, collecting and making judgements about evidence related to the goals of the learning being assessed. There is a wide range of ways of gathering evidence; which 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.

Popham's definition could apply to formative or summative purpose but assumes an active role of the teacher in the process: educational assessment refers to 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 National Research Council (NRC), 2001, p 20).

1.3.2 Summative assessment

Summative assessment refers to an assessment with a particular purpose: that of providing a record of a pupil's overall achievement in a specific area of learning at a certain time. It is a purpose that distinguishes it from assessment described as formative, diagnostic or evaluative (Department of Education and Science (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.

1.3.3 Teachers' assessment

Although teachers inevitably have a role in any assessment, the term 'assessment by teachers' (teachers' assessment, abbreviated to TA) is used for assessment where the professional judgement of teachers has a significant role in drawing inferences and making judgements of evidence as well as in gathering evidence for assessment. They may use observation during regular activities, or set up special tasks or projects to check what pupils can do or what ideas they have, or use class work (or course work), 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 that will frame and focus their attention, and in more formal approaches they may be using criteria or even checklists developed by others.

In some school-based assessment, teachers have a role only in gathering evidence that is then marked or graded by others. Since it 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 pupils learn. In summative assessment, this use is not a requirement since the purpose is primarily to report on learning to the various stakeholders – pupils, parents, other teachers, employers, assessment agencies, etc.

1.3.4 Other terms uses to describe assessment by teachers

Performance assessment is a term, used mainly in the United States, 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 include problem-solving, practical tasks, projects, extended writing, and

assessment methods such as interviews, observation, presentations, etc. 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' (p 1). *Embedded assessment* is a term that implies that, not only are the assessment tasks very similar to regular learning activities, but 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 takes 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 fall into the definition of TA used here.

All these types of TA combine some degree of formative purpose with summative purposes. 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 *on-going assessment* are ambiguous terms used sometimes 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.

1.4 The policy background

All teachers who report to their pupils' 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 pupils. However, these records serve what might be called 'internal' purposes of assessment, that is, they are for the information of those primarily concerned to help the further learning of the pupils. 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 pupils' 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 the internal.

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 pupils' achievements to parents at least once every year. For non-end-of-Key-Stage years, this is on the basis of the TA, although there are some optional, internally marked, tests which

teachers can use to help their judgments 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 TA is likely to have an impact on pupils, teachers or the curriculum depends on the degree of formality, the required mode of reporting or recording judgments, 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.

A different situation arises, generally with older pupils, where 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 limiting the proportion of credit 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; 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 pupils' motivation for learning (e.g. ARG, 2002; Harlen and Deakin Crick, 2003b), on pupils (Abbott *et al*., 1994; Pollard *et al*., 2000) and on the curriculum (Crooks, 1988; James, 1998; Shorrocks *et al*., 1993; 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), is giving rise to consideration of alternatives to testing. Thus, the Chief Inspector 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 pupils in their own school, and, more specifically, what strategies they will deploy to improve attainment' (Bell, 2003). The Qualifications and Curriculum Agency (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 that:

> 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.

1.5 The practice background

Despite the fact that conducting and reporting their summative assessment of pupils 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, stated the following:

> Where school-based teacher assessment is to be used for summative purposes then the relationship between teacher and pupil 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 teacher's 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 pupils against national curriculum attainment targets, or, later, against level descriptions and arrive at a 'level'. Before 1993, the 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 NC levels, which they grouped into three models of TA (Gipps *et al.,* 1995; McCallum *et al*., 1993). 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 headteachers judged this impact to be positive, making teaching more focused and teachers more aware of what children 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 pupils' achievement.

The Primary Assessment, Curriculum and Experience (PACE) project looked at the impact of the introduction of NCA on pupils. This project 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).

The report's authors also concluded that 'for these children, 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 pupils. In cases where the TA is all or part of assessment for an external award – high stakes for the pupil – 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, which has been in place in Queensland in some form since 1971. At first it was a norm-based assessment, but was converted to a criterion-based scheme in 1981. What makes this different from TA, which 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 program 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), posed the question, 'What is needed to make such an approach successful?', and answered 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 (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.

1.6 The research background

This review is closely related to the last 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 seven-year-olds in England and Wales caused teachers to plan in greater depth, although it had a less positive impact in increasing concentration on curriculum coverage at the expense of following their own or pupils' interests. Radnor (1996), Shorrocks *et al.* (1993), 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 (1997-1998) 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, although 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 programme (MASAP), Goldberg and Roswell (1999-2000) 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' 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.

1.7 Aims of the review and review questions

1.7.1 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 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.

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.

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.

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 of impact in particular circumstances will be sought so that, where trustworthy evidence is found, implications for policy and practice can be identified.

The outcomes will be 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, through a process of consultation with users, 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 pupils
- publication of the full report and of short summaries for different user groups in REEL
- identification of further research that is needed in this area

1.7.2 Review questions

Thus the main review question is:
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 will address the subsidiary question:
What conditions and contexts affect the nature and extent of the impact of using teachers' assessment for summative purposes?

The findings will be used to address the further question:
What are the implications of the findings for policy and practice in summative assessment?

1.7.3 Scope of the review

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

It is anticipated that there will be some studies in which the data reported are qualitative and may take the form of case studies. Others are likely to report statistical or judgemental evidence of impact made by teachers. All these will be included and a map of the types, designs and topic focus of studies will be created as part of the review.

2. METHODS USED IN THE REVIEW

2.1 User involvement

The users of this review will be all those involved with education; however, the review is concerned with matters relating to the use of assessment by teachers that influence decisions about policy. Thus the main focus is to inform policy-makers concerned with assessment, both at national and local levels, and practitioners and their professional bodies. The direct involvement of users in the conduct of the review will be through membership of the Review Group. The ALRSG includes the following users: a deputy secondary school headteacher 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 the Association of Assessment Inspectors and Advisers (AAIA), another is leading the review of assessment in Wales, and another is Director of the Learning to Learn project of the Economic and Social Research Council's (ESRC) Teaching and Learning Research programme. Eight members of the Review Group are members of the Assessment Reform Group and, through this, the Review Group has an ongoing relationship with the DfES, in particular with staff in charge of the Primary Strategy and the Key Stage 3 strategy.

Members of the Review Group will be involved in the process of the review by providing advice at meetings of the group, and between meetings by email, by providing information about studies through personal contacts, by taking part in keywording and data-extraction, and reflecting users' views in identifying implications of the findings.

2.2 Identifying and describing studies

2.2.1 Criteria for including and excluding studies

Inclusion criteria

The search for and selection of studies will be guided by the following inclusion criteria.

Language of the report: Studies included will, in general, be written in English. If papers in other European languages are found, arrangements will be made for translation. However, databases and journals primarily in languages other than English will not be searched.

Types of assessment: Studies will be included which deal with some form of summative assessment that is conducted by teachers. Studies reporting on purely formative assessment by teachers will not be included, but those where the assessment is for both formative and summative purposes will be included.

Study population and setting: Studies will be included which deal with assessment procedures and instruments used by teachers for assessing students, aged 4 to 18, in school.

Study type and study design: Studies will be included if they report information about changes in students, teachers or the curriculum that can be ascribed to assessment by teachers being used for summative assessment. Both naturally-occurring and researcher-manipulated evaluation study types will be relevant. Designs may include comparison of the experience of comparable classes with different experiences of TA for summative assessment or comparison of the same groups before and after the introduction of summative assessment based on TA. They may also include surveys of students' and teachers' perceptions of the impact of using TA for summative purposes and case studies of situations in which teachers' assessment is used for these purposes.

Topic focus: Since teachers' assessment can be used in all subjects, studies from all curriculum areas will be included. These will include both 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.

Studies meeting these criteria will be included in the descriptive map. If the number of such studies proves to be too large to take forward into the data-extraction stage, a narrower set of studies will be selected for in-depth review. The precise nature of these narrower criteria will be decided by the Review Group on the basis of the map. One possibility is that there could, for instance, be a restriction according to the purpose of the assessment, such as for certification or for national testing.

Exclusion criteria

Studies meeting the above inclusion criteria will be excluded for the following reasons and labelled accordingly:

A. Not summative assessment (exclude if information is gathered for formative purposes only; also exclude aptitude tests, special needs assessment)
B. Not assessment by teachers (exclude if assessment of teachers or studies of school evaluation; also exclude teacher-administered tasks or portfolios that are graded externally)
C. Not related to education in school (exclude studies relating to college students, higher education, nursing education, other vocational)
D. Not reporting impact of the process of assessment on students, teachers or the curriculum (exclude if the impact reported is a result of the outcome of the assessment and not the process)
E. Not research (exclude if not empirical study of particular procedures of assessment by teachers; also exclude reports of instrument development or description without report of use, or handbooks and reviews)

2.2.2 Methods for identifying studies

Studies will be identified from the following sources:

- Bibliographic databases (Educational Resources Information Center (ERIC), British Education Index (BEI), PsycLIT, Social Science Citation Index, Bath Information and Data Services (BIDS))
- Specialist registers (research registers of the National Foundation for Educational Research (NFER), the Scottish Council for Research in Education (SCRE), Center for Research on Evaluation, Standards, and Student Testing (CRESST))
- Search of journal publishers' web pages or handsearching of key journals
- Citation searches of key authors/papers
- Reference lists of key authors/papers

- References on key websites (Association for Educational Assessment (AEA) Europe, AAIA, NFER, QCA, Examination Boards)
- Personal contacts

The search for bibliographic databases will be for combination of the following terms:

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

Searches of the sources will be limited so as to identify studies conducted in the time period 1985–2003. This is the period of most relevance to current policy and practice in summative assessment; studies of practices before 1985 are likely to have been overtaken by the reforms in education in the later 1980s and 1990s, for example, the General Certificate of Secondary Education (GCSE) and the NCA in England and Wales. The review team will set up a database system using EndNote for keeping track of, and for coding, studies found during the review. Some titles and abstracts will be imported from databases (two-stage screening) and others entered manually into the first of these databases as a result of handsearching (one-stage screening).

2.2.3 Screening studies: applying inclusions and exclusion criteria

Full reports will be sought for those studies which appear, from abstracts, to meet the criteria or where there is insufficient information to be sure. The included reports will be collected into a second database. The inclusion and exclusion criteria will be re-applied to the full reports, during the course of keywording, and those not meeting these criteria will be excluded.

2.2.4. Methods for characterising included studies: keywording

The included studies will then be keyworded, using the EPPI-Centre's *Core Keywording Strategy: Data Collection for a Register of Educational Research*

(2002a). All these keywords are pre-fixed A. Additional keywords (prefixed B) which are specific to the context of the review (e.g. relating to types of impact reported in particular contexts) will be added to those of the EPPI-Centre. The review-specific keywords are as follows.

Review-specific keywords

B 1 Object of impact reported (not mutually exclusive)
- B. 1.1 On students
- B. 1.2 On teachers/teaching
- B. 1.3 On 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 Tasks set by teacher
- B. 3.4 Project
- B. 3.5 Regular work

B. 4 Type of assessment task (not mutually exclusive)
- B. 4.1 Special activity (timed or not timed)
- B. 4.2 Embedded (include regular work, but not portfolio)
- B. 4.3 Portfolio
- B. 4.4 Other

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
- B. 5.4 External for certification

All the keyworded studies will eventually be added to the EPPI-Centre's REEL, accessed via the EPPI-Centre website.

2.2.5 Quality assurance

Application of the inclusion and exclusion criteria and the keywording will be conducted by pairs of Review Group members and research assistants working, first independently, and then comparing their decisions and coming to a consensus. EPPI-Centre staff will also carry out a quality assurance role in applying inclusion and exclusion criteria and keywording for a sample of studies.

2.3 In-depth review

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

The descriptive map, based on the keywords, will be presented to the Review Group which will assist in identifying criteria for selection of those studies to be included in the in-depth review in the event that it is not appropriate or possible to extract data from all the mapped studies.

2.3.2 Method for extracting data from the studies

Studies identified for in-depth review will be analysed using the EPPI-Centre's detailed data-extraction *Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research* (EPPI-Centre, 2002b). Additional questions, which are specific to the context of the review, will be added to those of the EPPI-Centre.

2.3.3 Method for appraising quality of and weight of evidence from the studies

The EPPI 'weight of evidence' criteria will be used to help in making explicit the process of apportioning weights to the evidence provided by selected studies. This will involve judgements about three aspect of each study (A, B, C):

A. the soundness of methodology (informed by responses to questions about methodological coherence during the data-extraction), based upon the reports of the studies only
B. the appropriateness of the research design and analysis for answering the review question
C. the relevance of the study focus (from the sample, measures, conceptual focus, context or other indicator of the focus of the study) for answering the review question

The judgments for these three aspects will the combined into an overall weight of evidence, using an explicit rationale.

2.3.4 Methods for synthesising the findings of included studies

The data from the studies which meet the quality criteria relating to appropriateness and methodology will then be synthesised, initially by bringing together under separate heading the findings for impact on students, teachers and the curriculum. Since studies reporting impact on the curriculum tend also to report impact on teachers, it is likely that these will be 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). Having considered the impact in relation to these uses within types of impact, it will be useful to review findings across types of impact to explore the possibility that the nature of the impact may be different according to the use made of the results of the assessment.

2.3.5 Quality assurance

Data-extraction and assessment of the weight of evidence brought by the study to address the review question will be conducted by pairs of RG members and research assistants working, first independently, and then comparing their decisions and coming to a consensus. EPPI-Centre staff will also carry out a quality assurance role in the data-extraction process for a sample of studies.

2.3.6 Identification of implications

The findings of the review will be presented to, and discussed by, representatives of policy-makers, practitioners, researchers and others involved in education, in order to identify and validate implications for policy and practice.

3. REFERENCES

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.

Bell D (2003) Reporting England. Speech to the City of York Council's annual education conference. February. Available online 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 (1998) *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.

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 (1997-1998) 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.

Department of Education and Science (DES) (1997) *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.

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 (1999-2000) 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.

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 (1996) Assessment styles in the home countries. In: Boyle B, Christie T (eds) *Issues in Setting Standards*: *Establishing Comparabilities.* London: Falmer Press, pages 12-24.

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

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

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. In: *REEL*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

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 (2002) 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 C (1993) Teacher assessment at Key Stage 1. *Research Papers in Education: Policy and Practice* **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.

Plake BS, Impara JC, Fager JJ (1992) Assessment competencies of teachers: a national survey. Paper presented at the meeting of the National Council on Measurement in Education. San Francisco: April.

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 (2002) *Modern Educational Measurement: Practical Guidelines for Educational Leaders.* Needham, MA: Allyn and Bacon.

Radnor HA (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, Staintone R, Ring, K (1993) *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 H-W (2002) Same assessment, different practice: professional consciousness as a determinant of teachers' practice in a school-based assessment scheme. *Assessment in Education* **9:** 97-118.