REVIEW

March 2004

A systematic review of the evidence of reliability and validity of assessment by teachers used 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

AUTHOR AND INSTITUTIONAL BASE

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
Tel: 0117 928 7129

REVIEW TEAM MEMBERSHIP

*Mr R 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 the Association for Achievement and Improvement through Assessment (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 P Rayner, Link Inspector for Primary Education, Nottinghamshire
*Dr Gordon Stobart, Institute of Education, University of London

* Members who were actively involved at certain parts of the review.

ADVISORY GROUP MEMBERSHIP

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, London

ACKNOWLEDGEMENTS

This review was carried out with funding from the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-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 is gratefully acknowledged: Dina Kiwan, Zoe Garrett, Rebecca Rees and James Thomas. The author would also like to thank Max Coates for help in keywording and extracting data from studies.

LIST OF ABREVIATIONS

AAIA	Association for Achievement and Improvement through Assessment
ACCAC	The curriculum and assessment authority in Wales (no direct correspondence in English with the acronym)
'A' level	External examination normally taken by 18 year-olds and commonly required for university entrance
APU	Assessment of Performance Unit
ARG	Assessment Reform Group
ATL	Association of Teachers and Lecturers
CBM	Curriculum-based measurement
CCEA	Council for the Curriculum Examination and Assessment (Northern Ireland)
CSE	Certificate of Secondary Education
CTBS	Comprehensive test of basic skills
DAT	Differential aptitude test
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
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. KS1: ages 5 – 7; KS2, ages 7 – 11; KS3, aged 11 – 13; KS4, ages 14 – 16
LD	Level description (specifically the description of a level in the National Curriculum)
LEA	Local Education Authority (term used in England)
NA	Numerical ability
NC	National Curriculum (in England). The curricula and their titles are different in Wales, Northern Ireland and Scotland.
NCA	National Curriculum Assessment
NFER	National Foundation for Educational Research in England and Wales (producer of tests)
NUT	National Union of Teachers

'O' level	National external examination usually taken by 16 year-olds in England and Wales
PAT	Progressive achievement test
QCA	Qualifications and Curriculum Authority, overseeing the curriculum and assessment in England
SAT	Standard assessment task/test
SCAA	Schools Curriculum and Assessment Authority (predecessor of QCA)
SD	Standard deviation
SEN	Special educational needs
SES	Socio-economic status
SRA	Science Research Associates
TA	Teachers' assessment (see page 12)
UCCA	University Central Council on Admissions
VR	Verbal reasoning
WRT	Word recognition test
WSS	Work sampling scheme
Y (1, 2, etc.)	Year 1, 2, etc. Refers to years of school in England and Wales (Y1 – 5 years, Y2 – 6 years, etc.)

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

TABLE OF CONTENTS

SUMMARY

Background

The reason for proposing this review resulted from the work of the Assessment Reform Group (ARG) over several years and the more recent reviews conducted by the Assessment and Learning Research Synthesis Group (ALRSG), whose members include all the members of ARG. The review of classroom assessment initiated by ARG, and carried out by Black and Wiliam (1998), indicated that assessment used for formative purposes benefits teaching and learning, and raises standards of student performance. However, the ALRSG review, *A systematic review of the impact of summative assessment and tests on students' motivation for learning*, showed that high stakes tests can have a negative impact on students' motivation for learning and on the curriculum and pedagogy. But, summative assessment is necessary and serves important purposes in providing information to summarise students' achievement and progress for their teachers, parents, the students themselves and others who need this information. To serve these purposes effectively, summative assessment should interfere as little as possible with teaching methods and the curriculum and, importantly, should reflect the full range of learning outcomes, particularly those needed for continued learning and for learning how to learn.

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. Although assessment by teachers is used as the main source of information in some national and state assessment systems, in other countries, it has the image of being unreliable and subject to bias. This review was undertaken to provide some research evidence about the dependability of summative assessment by teachers and the conditions which affect it.

Definition of terms

Assessment is a term that covers any activity in which evidence of learning is collected in a planned and systematic way to draw inferences about learning. The purpose of the assessment determines how the information is used. Thus *assessment by teachers for summative purposes* means:

> 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

The phrase 'about their students' learning' excludes from this definition the role of teachers as markers or examiners in the context of external examination, where they do not mark their own students' work. It includes teachers' assessments of their own students as part of an examination for external certification. The phrase 'based on

their professional judgement' excludes assessment where information is gathered by teachers but marked externally, but would include students' self-assessment managed by teachers.

Reliability refers to how accurate the assessment is (as a measurement); that is, if repeated, how far the second result would agree with the first.

Validity refers to how well what is assessed matches what it is intended to assess. Different forms of validity derive from different ways of estimating it. Construct validity is a useful overarching concept.

Since reliability and validity are not independent of each other - and increasing one tends to decrease the other - it is useful in some contexts to refer to *dependability* as a combination of the two. The approach to summative assessment by teachers giving the most dependable result would protect construct validity, while optimising reliability.

Aims of the review

The aims of this review were as follows:

- to conduct a systematic review of research evidence to identify and summarise evidence relating to the reliability and validity of the use of teachers' assessment for summative purposes;
- to determine the conditions that affect the reliability and validity of teachers' summative assessment;
- to map the characteristics of studies reporting on the reliability and validity of teachers' assessment;
- in consultation with potential users of the review, to draw from this evidence implications of the findings for different user groups, including practitioners, policy-makers, those involved in teacher education and professional development, employers, parents and pupils;
- the identification of further research that is needed in this area and of the focus of subsequent reviews that might be undertaken by the ALRSG;
- publication of the full report and of short summaries for different user groups in the Research Evidence and Education Library (REEL).

Review questions

Thus the review was designed to address the main question:

What is the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?

and the subsidiary question:

What conditions affect the reliability and validity of teachers' summative assessment?

In order to achieve all the aims of the review, it was necessary to address the further question:

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

Methods

The review methodology followed the procedures devised by the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), and 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 validity and/or reliability of methods used. Bibliographic databases and registers of educational research were searched online as were relevant online journals, with other journals and back numbers of those only recently put 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 and 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 *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 independently by two reviewers who then worked together to reach a consensus, using the EPPI-Reviewer (*Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research* (EPPI-Centre, 2002b)). Review-specific questions relating to the weight of evidence of each study in the context of the review were used in addition to those of the EPPI-Reviewer. 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 for the synthesis of evidence from the in-depth review was based on the extent to which the studies were concerned with reliability or validity of the assessment. Despite the difficulty in making a clear distinction between these concepts, and their inevitable interdependence, it was possible to designate each one as providing evidence *primarily* in relation to reliability or *primarily* in relation to validity. Evidence in relation to the conditions affecting reliability or validity was drawn together separately. In the synthesis and discussion, reference was made to the weight of evidence provided by each study.

Potential users of the review were involved in several ways: providing advice as members of the review group; providing information about studies through personal contact; participating in keywording and in data-extraction; and through a consultation seminar on implications of the draft findings of the review attended by a number of policy and practitioner users.

Results

Identification of studies

The search resulted in a total of 431 papers being found. Of these, 369 were excluded, using exclusion criteria. Full texts were obtained for 48 of the remaining 62 papers, from which a further 15 were excluded, and two sets of papers (three in one case and two in the other) were linked as they reported on the same study. This left 30 studies after keywording. All of these were included in the in-depth review.

Systematic map

The 30 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: 15 were conducted in England, 12 in the United States and one each in Australia, Greece and Israel. All studies were concerned with students between the ages of 4 and 18. Of the 30, 11involved primary school or nursery students (aged 10 or below) only, 13 involved secondary students (aged 11 or above) only, and six were concerned with both primary and secondary students. There was no variation across educational settings in terms of whether the study focus was on reliability or validity, but there were slightly more evaluations of naturally-occurring situations in primary schools. Almost all studies set in primary and nursery schools involved assessment of mathematics and a high proportion related to reading. At the secondary level, studies of assessment of mathematics and 'other' subjects (variously concerned with foreign languages, history, geography, Latin and bible studies) predominated.

Eighteen 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 nine were either set externally or set by the teacher to external criteria. The vast majority were assessed by teachers, using external criteria. The most common purpose of the assessment in the studies was for national or state-wide assessment programmes, with six studies related to certification and another six to informing parents (in combination with other purposes). As might be expected in the context of summative assessment, most research related to the use of external criteria by teachers, with little research on student self-assessment or teachers using their own criteria.

In-depth review

Findings from studies of reliability of assessment based on teachers' judgements

There was evidence of high weight for the following:

- the reliability of portfolio assessment where tasks were not closely specified was low (Koretz *et al.,* 1994; Shapley and Bush, 1999); this finding has been used as an argument for increasing the match between task and assessment criteria by closer specification of tasks.
- The finer specification of criteria, describing progressive levels of competency, has been shown to be capable of supporting reliable teachers' assessment (TA)

while allowing evidence to be used from the full range of classroom work (Rowe and Hill, 1996).
- Studies of the National Curriculum Assessment (NCA) for students aged 6 and 7 in England and Wales in the early 1990s, found considerable error and evidence of bias in relation to different groups of students (Shorrocks *et al.*, 1993; Thomas *et al.,* 1998).
- Study of the NCA for 11 year-olds in England and Wales in the later 1990s shows that results of TA and standard tasks agree, and are to an extent consistent with the recognition that they assess similar but not identical achievements (Reeves *et al.,* 2001).
- The clearer teachers are about the goals of students' work, the more consistently they apply assessment criteria (Hargreaves *et al.*, 1996).
- When rating students' oral proficiency in a foreign language, teachers are consistently more lenient than moderators, but are able to place students in the same rank order as experienced examiners (Good, 1988a; Levine *et al.*, 1987).

There was evidence of medium weight for the following:
- Interpretation of correlations of TA and standard task results for seven year-olds should take into account the variability in the administration of standard tasks (Abbott *et al.*, 1994).
- Teachers who have participated in developing criteria are able to use them reliably in rating students' work (Hargreaves *et al.*, 1996; Frederiksen and White, 2004).
- Teachers are able to score hands-on science investigations and projects with high reliability using detailed scoring criteria (Frederiksen and White, 2004; Shavelson *et al.,* 1992).

Findings from studies reporting the validity of assessments based on teachers' judgement

There was evidence of high weight for the following:
- Teachers' judgement of the academic performance of young children are influenced by the teachers' assessment of their behaviour; this adversely affects the assessment of boys compared with girls (Bennett *et al.*, 1993).
- The introduction of TA as part of the national curriculum assessment initially had a beneficial effect on teachers' planning and was integrated into teaching (Hall *et al.*, 1997); subsequently, however, in the later 1990s, there was a decline in earlier collaboration among teachers and sharing interpretations of criteria, as support for TA declined and the focus changed to other initiatives (Hall and Harding, 2002).
- The validity of a science project as part of 'A' level examinations for assessing skills different from those used in regular laboratory work was reduced when the project assessment was changed from external to internal by teachers (Brown, 1998).
- Teachers judgements guided by checklists and other materials in the Work Sampling System were found to have high concurrent validity for assessment of kindergarten (Kg) to Grade 3 students (Meisels *et al.*, 2001).
- Teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching (Coladarci, 1986).

There was evidence of medium weight for the following:

- There is variation of practice among teachers in their approaches to TA, type of information used and application of national criteria (Gipps *et al.*, 1996; Radnor, 1995).
- There is conflicting evidence as to the relationship between teachers' ratings of students' achievement and standardised test score of the same achievement when the ratings are not based on specific criteria (Hopkins *et al.*, 1985; Sharpley and Edgar, 1986).
- The rate at which young children can read aloud is a valid curriculum-based measure of reading progress as measured by a standardised reading test (Crawford *et al.*, 2001).
- Tentative estimates of construct validity of portfolio assessment, derived from evidence of correlations of portfolios and tests, were low (Koretz *et al.*, 1994; Shapley and Bush, 1999).
- Teacher assessment of practical skills in science makes a valid contribution to assessment at 'A' level within each science subject, but there is little evidence of generalisability of skills across subjects (Brown *et al.*, 1996).
- Teachers' perceptions of students' ability and probability of success on a test are moderately valid predictors of performance on the test, as are student self-assessments of their performance on a test after they have taken it (Wilson and Wright, 1993).

Evidence in relation to the conditions that affect the reliability and validity of teachers' summative assessment

Both high and medium weight evidence indicated thefollowing:

- There is bias in teachers' assessment (TA) relating to student characteristics, including behaviour (for young children), gender and special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills.
- There is variation in the level of TA and in the difference between TA and standard tests or tasks that is related to the school. The evidence is conflicting as to whether this is increasing or decreasing over time. There are differences among schools and teachers in approaches to conducting TA.
- There is no clear view of how reliability and validity of TA varies with the subject assessed. Differences between subjects in how TA compares with standard tasks or examinations results have been found, but there is no consistent pattern suggesting that assessment in one subject is more or less reliable than in another.
- It is important for teachers to follow agreed procedures if TA is to be sufficiently dependable to serve summative purposes. To increase reliability, there is a tension between closer specification of the task and of the conditions under which it is carried out, and the closer specification of the criteria for judging performance.
- The training required for teachers to improve the reliability of their assessment should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of the language used. Training should also focus on the sources of potential bias that have been revealed by research.
- Teachers can predict with some accuracy their students' success on specific test items and on examinations (for 16 year-olds), given specimen questions. There is less accuracy in predicting 'A' level grades (for 18 year-olds).

- Detailed criteria describing levels of progress in various aspects of achievement enable teachers to assess students reliably on the basis of regular classroom work
- Moderation through professional collaboration is of benefit to teaching and learning as well as to assessment. Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others, including assessment advisers can give.

Conclusions

The implications of the findings of the review were explored through consultation with invited teachers, head teachers, researchers, representatives of teachers' organisations, of the Association for Achievement and Improvement through Assessment (AAIA), and of UK government agencies involved in national assessment programmes. Some points went beyond the review findings and are listed separately after those directly arising from the research evidence.

Implications for policy

(a) When deciding the method, or combination of methods, of assessment for summative assessment, the shortcomings of external examinations and national tests need to be borne in mind.
(b) The essential and important differences between TA and tests should be recognised by ceasing to judge TA in terms of how well it agrees with test scores.
(c) There is a need for resources to be put into identifying detailed criteria that are linked to learning goals, not specially devised assessment tasks. This will support teachers' understanding of the learning goals and may make it possible to equate the curriculum with assessment tasks.
(d) It is important to provide professional development for teachers in undertaking assessment for different purposes that address the known shortcomings of TA.
(e) The process of moderation should be seen as an important means of developing teachers' understanding of learning goals and related assessment criteria.

Implications for practice

(a) Teachers should not judge the accuracy of their assessments by how far they correspond with test results, but by how far they reflect the learning goals.
(b) There should be wider recognition that clarity about learning goals is needed for dependable assessment by teachers.
(c) Teachers should be made aware of the sources of bias in their assessments, including the 'halo' effect, and school assessment procedures should include steps that guard against such unfairness.
(d) Schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers is sustained: for example, by protecting time for planning assessment, in-school moderation, etc.
(e) Schools should develop an 'assessment culture' in which assessment is discussed constructively and positively, and not seen as a necessary chore (or evil).

Implications for research

(a) There should be more studies of how teachers go about assessment for different purposes, what evidence they use, how they interpret it, etc.

(b) The reasons for teachers' over-estimation of performance, compared with moderators' judgements of the same performance, need to be investigated to find out, for instance, whether a wider range of evidence is used by the students' own teachers, or whether criteria are differently interpreted.
(c) More needs to be known about how differences between schools influence the practice and dependability of individual teachers.
(d) Since evaluating TA by correlation with test results is based on the false premise that they assess the same things, other ways need to be found for evaluating the dependability of TA.
(e) There needs to be research into the effectiveness of different approaches to improving the dependability of TA, including moderation procedures.
(f) Research should bring together knowledge of curriculum planners, learning psychologists, assessment specialists and practitioners to produce more detailed criteria that can guide TA.

Additional points related to the review identified in consultation with users

(a) It is important to consider the purpose of assessment in deciding the strengths and weaknesses of using teachers' assessment in a particular case. For instance, when assessment is fully under the control of the school and is used for informing pupils and parents of progress ('internal purposes'), the need to combine TA with other evidence (e.g. tests) may be less than when the assessment results are used for 'external' purposes, such as accountability or the school or selection or certification of students.
(b) There needs to be greater recognition of the difference between purposes of summative assessment and of how to match the way it is conducted with its purpose. For instance, the 'internal' assessment that is under the control of the school should not emulate the 'external' assessment which has different purposes.
(c) If tests are used, they should be reported separately from TA, which should be independent of the test scores.
(d) There is evidence that a change in national assessment policy is due. The current system is not achieving its purpose. The recent report on comparability of national tests over time (Massey *et al.*, 2003) concludes that TAs have shown less change in standards than the national tests. The authors state, 'National testing in its current form is expensive, primarily because of the external marking of the tests, and the time may soon come when it is thought that these resources may make a better contribution elsewhere' (Massey *et al.*, 2003, p 239).
(e) Improving teachers' formative assessment would also improve their summative assessment and so should be a part of a programme of professional development aimed at enabling teachers' judgements to be used for summative purposes.
(f) The role that pupils can take in their own summative assessment needs to be investigated and developed.
(g) Any change towards greater use of TA in current systems where summative assessment is dominated by tests requires a major switch in resources from test development to supporting teacher-led assessment.
(h) Change towards greater use of TA for summative purposes, requires a long-term strategy, with strong 'bottom-up' elements and provision for local transformations.

1. BACKGROUND

This chapter begins by summarising the events leading to the proposal for this systematic review, the third conducted by the Assessment and Learning Research Synthesis Group (ALRSG). The review takes place at a time when there is interest at the highest levels of government agencies concerned with the school curriculum and assessment in giving a greater role to assessment by teachers in summative assessment. The discussion of the meaning of terms, such as *summative assessment*, *reliability* and *validity*, in the second section, is followed by some expansion on the policy, practice and research background. The chapter concludes with a statement of the review questions.

1.1 Aims and rationale

A previous review conducted by the Assessment and Learning Research Synthesis Group (ALRSG): *A systematic review of the impact of summative assessment and tests on students' motivation for learning*, showed that tests can have a negative impact on students' motivation for learning and on the curriculum and pedagogy. An important implication drawn from the review was to call into question the validity of tests; it also raised some ethical issues, given the differential impact of tests on low achieving students. Nevertheless, summative assessment is necessary and serves an important purpose in providing information to summarise students' achievement and progress for their teachers, parents, the students themselves and others who need this information.

For summative assessment to serve its purposes effectively, without distorting teaching methods and the curriculum, it needs to reflect the range of learning outcomes that are important aims in the 'information age' – in particular, learning to learn and motivation for continued learning throughout life (for example, Organisation for Economic Co-operation and Development (OECD), 1999, 2001) and other educational goals that are not readily amenable to formal testing. It also needs to take a form that benefits all students. The use of the information that teachers can gather through their constant contact with students has the potential to do this. As part of their regular work, teachers can build up a picture of students' attainments across the full range of activities and goals. This gives a broader and fuller account than can be obtained through any test using a necessarily restricted range of items and so can be described as a more valid means of assessing outcomes of education (Crooks, 1988; Wood, 1991). Further, in this process, the teacher has the opportunity to use this accumulating information to help learning.

Assessment conducted by teachers is variously called ongoing, continuous, school-based or, in the UK, teacher assessment. The last of these is somewhat confusing, suggesting assessment of teachers rather than by teachers. To avoid confusion, in this documents it is referred to as assessment by teachers (TA). TA can serve formative purposes (assessment *for* learning) and summative purposes (assessment *of* learning). The distinction lies not in how the information is gathered but in how it is used, as discussed further in section 1.2. This study focuses on the summative use,

for reporting on learning rather than the formative use, for helping learning, although use for the latter purpose does not exclude use for the former also.

TA can take a range of forms, such as using prescribed tasks which are administered and marked by teachers, teacher-made tests, or the use of a set of criteria in relation to regular class work and observation of learning processes. In all cases the summative use of the assessment means that the question of whether teachers are using a mark scheme or criteria in comparable ways has to be addressed. Thus one of the principal issues in using TA relates to reliability. Where summative assessment is exclusively based on teachers' assessment, as in the state of Queensland, Australia, there are moderation procedures that are designed to ensure comparability of standards (Maxwell, 1995). Moderation, or quality assurance, can be conducted in various ways, such as through the use of detailed criteria or descriptors, through holding moderation meetings within and across schools, providing exemplars of the application of criteria in practice, external moderation or inspection (Harlen, 1994). The extent to which these practices are effective, in what conditions, is important information for informing decisions about using TA for summative purposes. Since reliability and validity are interconnected concepts, however, it is necessary to consider both and it is the aim of this review to do so.

From this brief view of some of the background to this review, it appears that there is a strong case for giving teachers' assessment a greater role in summative assessment. However, implementation is not without considerable problems and more information is needed about teachers' summative assessment in practice and the conditions which are associated with acceptable levels of reliability and validity.

Thus the aims of this review are: to identify and summarise evidence relating to the reliability and validity of the use of TA for summative purposes and to the conditions that affect the reliability and validity of teachers' summative assessment; and to draw from this evidence implications for policy and practice in summative assessment.

1.2 Definitional and conceptual issues

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 and which is chosen depends 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.

Consider a widely quoted definition by Popham:

> 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).

This could apply to formative or summative purpose but assumes an active role of the teacher in the process.

Summative assessment

The term '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 the 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.

Teachers' assessment

Although teachers inevitably have a role in any assessment, the term 'assessment by teachers' (teachers' assessment or 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. Teachers 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 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, while externally produced tests or other assessment procedures serve a summative function. However, this is not by any means always the case. While 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.

Reliability

Reliability of the result of an assessment, which may be in the form of a test score or summary grade or mark, is the extent to which it can be said to be accurate and not influenced by, for instance, the particular occasion or who does the marking or grading. Thus reliability is often identified as, and measured by, the extent to which,

'if the assessment were to be repeated, the second result would agree with the first' (Harlen, 2000, p 111). When it is not possible to give the same test twice to the same pupils, or to repeat observations of a particular event assessed by teachers, other procedures are adopted. In the case of tests, these include using parallel forms of the test, or splitting the test randomly into two halves and comparing the scores or correlating the items with the total score and averaging the result. In the case of observation of tasks, the equivalent procedures are to compare the rating of the same event by two independent raters.

These approaches, however, are based on the assumption that there is a 'true measure' to be found, an assumption which is increasingly challenged. There are many different reasons for variation between one occasion and another and it is recognised that there is no possibility of 100% freedom from error in an assessment. Moreover, the concept of reliability as the 'error' in measurement is questioned when the range of the assessment is widened and the situations in which performance is assessed are not standardised; these are features of TA. Indeed Gipps (1994, p 171) suggests dropping the term *reliability* for assessment other than standardised tests, in favour of *comparability*. She suggests that this change is part of the paradigm shift from a 'testing' model, based on psychometrics and an assumption that there is a 'true score', to a broader model of 'educational assessment' which recognises that assessment is not an exact science.

Validity

Validity refers to what is assessed and how well this corresponds with the behaviour or construct that it is intended to test or assess. The distinction between reliability and validity is clear in, for example, the case of a multiple-choice test of knowledge about materials that conduct electricity. This would not be a valid assessment of understanding of an electric circuit, although it would give a score of quite high reliability (Harlen, 2000). Validity, however, is not a simple concept and various forms of it are identified according to the basis of the judgement of validity.

Evidence relating to the content validity of an assessment would result from comparing the content assessed with the content of a curriculum it was intended to assess. Face validity is based on expert judgement of what an assessment appears to assess, while predictive validity is the extent to which an assessment reflects an intended future performance. Concurrent validity is derived from the correlation of the outcomes of one assessment procedure with another that is assumed to assess the same knowledge or skill. Construct validity is a judgement of how well the assessment calls upon the knowledge and skills or other constructs that are supposedly assessed; it requires a clear definition of the domain being assessed, and evidence that, in the assessment process, the intended skills and knowledge are used by the learners.

A further form of validity of increasing interest and relevance is consequential validity, articulated by Messick (1989). He proposed that 'what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that interpretation entails' (p 9). In other words, the uses of and consequences of the uses of a test are what determine its validity. If inappropriate use is made of tests which make them unfair in an ethical and social sense, then however *technically* valid, the tests lack *consequential* validity.

Moreover, if no use is made of the results of a test, then it, too, lacks consequential validity.

This plethora of different kinds of validity – and the above is by no means a complete list – has led to a call for a unified view of validity in terms of an overarching concept which subsumes several of the separate types (e.g. Black, 1998; Gipps, 1994). James (1998), for example, reports a view held by many that construct validity is the unifying concept.

The relationship between reliability and validity

It is well recognised that the concepts of reliability and validity are not independent of each other in practice. The relationship is usually expressed in a way that makes reliability the prior requirement. The argument is that an assessment that does not have high reliability cannot have high validity; if there is uncertainty about the accuracy of the assessment and it is influenced by a number of different factors, then the extent to which it measures what it is intended to measure must also be uncertain. However, this argument tends to lead to attempts to increase reliability which generally means closer and closer specification, and use of methods that have the least error. It results in gathering and using a restricted range of evidence, leading to a reduction in validity. On the other hand, if validity is increased by extending the range of the assessment to include outcomes such as higher level thinking skills, then reliability is likely to fall, since these aspects of attainment are not easily assessed. However, while this may mean that, for the summative assessment, there has to be a compromise between reliability and validity, when the data are used for formative assessment, validity is paramount and reliability of less importance. As Sadler (1989, p 122) points out, ‘Attention to the validity of judgements about individual pieces of work should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement. Reliability will follow as a corollary’. Thus the relationship is a particular issue if an attempt is made to use the same assessment information for both formative and summative purposes.

Dependability

The recognition of the interaction between validity and reliability means that, while it is useful to consider each separately, what matters in practice is the way in which they are combined. This has led to the combination of the two in the concept of dependability (Wiliam, 1993). James (1998, p 159) expresses this as:

Reliability + Validity = Dependability

However, there is no simple sum to be calculated here. Since, as noted above, it is not possible to have high reliability and high validity, it is necessary to consider the balance of priorities. In deciding the relative importance of the two components of dependability, the purpose of the assessment has to be taken into account. Thus, in the case of TA for summative purposes, where the reason for adopting this approach rather than using tests is to protect construct validity, it is important to consider what is the highest optimum reliability that can be reached while preserving construct validity for the products of the assessment to serve its purpose. This would identify the approach which gives the most dependable assessment.

1.3 Policy and practice background

1.3.1 Policy

TA has come to be associated more with a formative role than a summative one but it 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 (Broadfoot *et al.*, 1990; Wood, 1991; Black, 1998). It is widely used as the only form of assessment for many post-graduate courses and for vocational and professional certification. At the school level, in England and Wales it became a part of the procedures for the Certificate of Secondary Education (CSE) introduced in the 1950s and was soon after taken up by some of the General Certificate of Education (GCE) examination boards. According to Black (1998, p 15), 'at least one [board] developed an experimental system in the 1960s which enabled pupils to gain a GCE in English entirely on the basis of teacher assessment with no external written examination'. When the dual CSE/GCE system of examinations was replaced by the General Certificate of Secondary Education (GCSE) in 1988 in England and Wales, assessment by teachers was incorporated 'in most subjects, covering those important aspects which by their nature could not be assessed by external written tests' (ibid). In England and Wales, the part played by TA was contentious and, in 1992, the proportion of credit awarded on the basis of teacher assessment in the GCSE was limited by government decree, reflecting some distrust of teachers' judgement. In Scotland, where the equivalent to GCSE, called the Standard Grade of the Scottish Certificate of Education, was introduced from 1984, there remains a considerable component of teacher-based assessment and the role of teachers' judgements continues to be a key one in all secondary school qualifications (Harlen, 1995).

Teachers' assessment was a central aspect of the Natiojnal Curriculum Assessment (NCA) of England and Wales introduced following the Task Group on Assessment and Testing (TGAT) report of 1987 (DES, 1987). One reason for this was to provide greater validity of assessment of certain outcomes for younger children, but it was also to allow information gathered by teachers to be used formatively as well as in combination with standard test results for summative purposes. While the Dearing Review recommended parity of esteem between teachers' assessment and national test results (James, 1998), in practice, tests take a stranglehold on the curriculum and teaching and have had a progressively more negative impact on pupil motivation for learning (Harlen and Deakin Crick, 2002).

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. There are problems, however, in ensuring that criteria are sufficiently explicit to support reliable summative data (e.g. Jones and Striven, 1991).

1.3.2 Practice

The use of assessment by teachers for summative purposes is often advocated on the grounds that:

(a) it reduces the pressure on teachers and students from external examinations;
(b) it enables teachers greater freedom to pursue and assess their own goals;

(c) it can be conducted as part of teaching and so provide formative feedback to students, thus improving their learning (Crooks, 1988).

However, experience in practice has often been disappointing in failing to live up to these claims. For instance, unless teachers are prepared for taking advantage of the autonomy that is theoretically available to them, the tendency is for them to interpret ongoing or continuous assessment as a series of tests. Although these are teacher-made, they tend to emulate the form and scope of external tests. This seems to be particularly so when the TA is a component of a summative assessment, with the remainder (often more than 50%) coming from an examination. Black (1993, p 83) notes that this arrangement, common in science education where practical work is teacher-assessed, may mean that 'prescriptions which are intended or perceived as rigid may simply convert the classroom into a formal examination room on special occasions'.

From more recent experience, Lubisi and Murphy (2002, p 264) report that, in the operation of continuous assessment in Natal, South Africa, 'there is ample evidence to suggest that the system was dominated by the use of (admittedly teacher-produced) tests which were modelled on the very matric exam which supposedly threatened teacher autonomy in the first place'. If teachers are given the role of assessor before being prepared for taking it, it is little wonder that the implementation falls far short of intentions.

The intention of encouraging teachers to give feedback and use their assessment formatively may also founder in a situation where, either through pressure of numbers in the class or established practice, the only feedback that the students are given is in the form of marks. Research by Butler (1988) shows that such feedback fails to have a positive impact on students' performance.

Reasons for the frequent failure of summative assessment by teachers to live up to the expectations for it are most often given in terms of their preparation for this part of their work. Many teachers have a narrow view of assessment and do not know how to respond to freedom to use evidence from students' actions, projects and processes. Merely being required to follow given criteria or guidelines is not enough. Even the discussion of students' work with colleagues, supposedly for the purpose of moderation, can become an exercise in adjusting marks. Donnelly *et al.* (1993), in a report on the internal assessment of practical work in science, found that the use of external assessors can lead to a loss of responsible autonomy, with teachers concerned about 'passing' the moderation.

The importance of thorough professional development is underlined by evidence from a study of the markers in the Maryland School Performance Assessment Program (MSPAP). This showed that being involved in marking experiences was not enough to influence their classroom practices in performance-based assessment nor their understanding of performance-based assessment (Goldberg and Roswell, 1999-2000). However, research by Gilmore (2002) in New Zealand, involving teachers who administered individual assessment tasks to children as well as marking a range of students' work, including video-taped performances, in the context of a national evaluation monitoring project (NEMP), reported a positive impact of these experiences. It was anticipated that this involvement would constitute useful professional development for the administrators and markers (Crooks and

Flockton, 1993) since the teachers were exposed to 'high quality assessment procedures'. Similar claims were also made of the involvement of teachers in the Assessment of Performance Unit's practical tests (Johnson, 1989) but were not formally investigated.

In the NEMP project, the teacher administrators received training before working 'intensely with 60 children in at least five different schools administering NEMP tasks' (Gilmore, 2002, p 349). The teacher markers were responsible for formulating the marking criteria for the assessment tasks and undertook the marking, working in pairs where substantial judgements was needed. Both assessment administrators and markers reported developing confidence in their knowledge and understanding of assessment, among other benefits. Gilmore (2002) points out the contrast between the experience of markers in the MSPAP and in NEMP. In the MSPAP 'teachers were trained to reach 70% exact agreement with pre-determined marking criteria; the marking was not carried out centrally, therefore there were limited opportunities for interaction between teachers; all children's performances were recorded in writing in individual test booklets, no 'live' performances captured on video placing an emphasis on the product of children's assessment rather than process' (p 347). These differences point to some of the features that seem to be required in teachers' experience in order to develop their understanding, and possibly their practice, of assessment.

Professional development is needed for a variety of purposes: to ensure that teachers have the skills, knowledge and support to conduct assessment effectively; to ensure quality control so that users can have confidence in the teachers' judgements; and to help teachers reconcile the dual role that they are required to take in both promoting and judging learning. The task is particularly difficult in countries where a great deal of emphasis is given to examinations results. For instance, Choi (1999) reporting on assessment in Hong Kong, points out that support of the teaching profession is paramount.

> 'A school-based assessment initiative …is doomed to failure if a top-down approach is adopted …To secure teachers' support, more assessment training and resource support for teachers are essential. Under a school-based assessment system, teachers are under pressure because they wear two hats, as facilitators of learning and as examiners. Where one role ends and the other begins could pose considerable problems, particularly for new teachers. The next difficulty is to ensure credibility for school-based assessment. The Authority needs an effective and efficient quality assurance and quality control system to assure the users of examination results, such as employers and tertiary institutions, as well as the general public, of the reliability of this scheme of assessment. This is not a simple task.' (Choi, 1999, p 415).

Such observations may seem to suggest that obstacles to introducing and supporting a system in which assessment by teachers is used summatively are almost overwhelming. However, as several supporters of using TA have pointed out (e.g. Black, 1986), attempts to develop and implement TA as part of assessment systems have not received even a small part of the resources currently used to running external examinations.

1.4 Research background

Satterley (1994) reviewed the evidence of the reliability of examinations and found that there was little to support the popular assumption that they have high values of technical reliability. Studies which compare TA with external examinations are therefore necessarily contaminated by the limited extent of the reliability of the external measures. To avoid this, many studies of reliability of TA focus on internal consistency of the teachers' judgements. For instance, Black (1993) quotes a study of Australian teachers' use of reading and writing scales, where Griffin (1989) found internal consistency was high (Cronbach alphas 0.88 and 0.87), although correlations with external tests of reading were between 0.55 and 0.72. Black (1993, p 55) reports:

> 'Griffin judged that this consistency was as good as can be achieved with external tests, and concluded that the correlations are as high as one would expect from internal assessments if they had reliability comparable with that of the external tests, given that the external tests were themselves of limited reliability'.

A meta-analysis by Hoge and Coladarci (1989) of studies of TA compared with achievement tests found a mean correlation of 0.69, with higher values for higher-achieving than for lower-achieving pupils. Even though there were differences between studies, they concluded that, overall, the teachers' judgements were of greater validity than the tests and that their findings indicated that TAs deserved to be given the same attention as other measures of achievement.

Some early studies concerned the extent to which teachers could reliably predict examination or test scores and were thus constrained by what these external instrument measured. Thus Murphy (1979) compared predicted grades for 'O' and 'A' level students and could come to no greater conclusion than that 'higher than expected grades were explained by better than expected performance in the examination, whereas lower that expected grades were explained either by worse than expected performances in the examination or by differences in standards between teachers and examiners' (Murphy 1979, p 54). Hoge and Butcher (1984) also investigated the accuracy with which teachers could estimate student performance in a standardised reading test that was familiar to them. In addition to reading achievement, the researchers asked for teachers' ratings of the students' general ability and motivation for school work. They found a high level of agreement between the actual and the predicted achievement test scores, but also reported that some teachers over-estimated the performance of high-ability students and under-estimated the performance of the lower-ability students. This showed in a strong correlation between the teachers' judgements of reading achievement and general ability.

A criticism frequently made of TAs is that they are subject to bias, according to factors such as gender and general ability, as just noted in the Hoge and Butcher (1984) study. Given that research (e.g. Spear, 1984) has shown that written work can be differently assessed by teachers according to whether the student is known to be male or female, the effect is considered likely to be greater when the assessment is carried out face to face. Wood (1991) included in his comprehensive review of

research on assessment and testing a number of studies of various forms of bias in TAs. Several studies showed that teachers were more likely to be biased by their opinion of students' work habits rather than their social behaviour.

The conditions influencing the practice of assessment by teachers include the way in which they interpret the requirements for summative assessment and its associated guidelines or regulations. Yung (2002) provides case study evidence that this is influenced by teachers' confidence and professional consciousness. In his study of the implementation of regulations for school-based assessment of practical work in biology, he found some teachers who followed the regulations mechanically and felt that this constrained their teaching, while others took advantage of the assessment to adapt teaching in a way that enhanced the students' learning.

The seriousness for individual pupils of the error in assessment, whether the assessment is based on a test or teachers' judgement, depends on the way in which the results are used. Black (1998, p 41) points out the 'Even with a reliability coefficient that is high enough to be commonly regarded as acceptable (say 0.85 to 0.9) the errors in pupils' scores that are implied may means that a significant proportion are given the wrong grade'. He also cites Wiliam (1995), who estimated that, based on the internal consistency of the Key Stage 3 national tests in England, 30% of pupils are likely to be placed at the wrong level. Similarly, at the boundary between passing and failing the 11+ in Northern Ireland, many pupils are divided by a small number of marks which are within the margin of error of the tests.

1.5 Authors, funders, and other users of the review

This review is the third EPPI review carried out by the 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 because of the recent interest in alternatives to testing for summative assessment, in the form of assessment by teachers. Recent evidence of this interest comes from the commissioning by Qualifications and Curriculum Authority (QCA) in England of 'Experiences of Summative Teacher Assessment in the UK', the establishment of the Daugherty Review of Assessment Group in Wales and the Assessment is for Learning project in Scotland.

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 –2005, on the topic of 'Assessment systems of the future: the place of assessment

by teachers'. The findings of this review will be the main input into the first seminar in the series, attended by policy-makers, advisers and teachers from all parts of the UK.

1.6 Review questions

The main review question was:

What is the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?

To achieve its aims the review addressed the subsidiary question:

What conditions affect the reliability and validity of teachers' summative assessment?

The findings are used to address the further question:

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

2. METHODS USED IN THE REVIEW

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 User involvement

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 dependability 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 is through membership of the Review Group. The ALRSG includes the following users: a deputy secondary 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 is leading 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. Seven (eight in January 2004) of the Review Group are members of the ARG and, through this, the Review Group has an ongoing relationship with the Department for Education and Skills (DfES), in particular with staff in charge of the Primary Strategy and the KS3 strategy.

2.1.2 Methods used

Users have been involved in the review in four ways:

1. As members of the Review Group, in attending regular meetings to advise at key points of the review and at other times through email. Four meetings were held in 2003.
2. Providing information about studies through personal contact.
3. Participating in keywording and in data-extraction; five members were actively involved in this way.
4. Involvement in setting up and conducting a consultation on implications of the draft findings of the review with a wider range of policy and practitioner users. This took the form of a two-day invitational seminar, held on January 12th and 13th 2004. The 24 participants included, in addition to the members of the ARG, senior staff of QCA, the curriculum and assessment authority in Wales (ACCAC), the Council for the Curriculum Examination and Assessment (Northern Ireland) (CCEA), the Association of Teachers and Lecturers (ATL), the National Union of

Teachers (NUT), AAIA, the Scottish Executive, and primary and secondary head teachers. The seminar participants discussed the issues surrounding the review, the findings and the implications for policy, practice and research (see section 5.4).

2.2 Identifying and describing studies

2.2.1 Defining relevant studies

The following criteria were drawn up in order to decide which of the studies were to be included in the review.

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
Studies were included if they reported information about the validity and/or reliability of methods used by teachers for summative assessment. 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 TA for summative purposes and case studies of TA used for these purposes.

Topic focus
Since TA 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.

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

2.2.2 Identification of potential studies: search strategy

Studies were identified through a combination of a two-stage strategy, used for

databases where there is no immediate screening, and a one-stage strategy, where handsearching allowed immediate screening.

The two-stage search was begun by searching bibliographic databases and registers of educational research. 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. Study titles and abstracts were reviewed in relation to the inclusion and exclusion criteria before being entered onto the database. Other studies were found by scanning the reference 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 included in an EndNote database, each being labelled with its source and method of identification.

2.2.3 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 crtieria. Each excluded study was labelled with the reasons(s) for exclusion. A sample of titles and abstracts (15%), including both those judged to be included and excluded, was reviewed by EPPI-Centre staff for the purpose of quality assurance.

2.2.4 Characterising included studies: keywording

The included studies were keyworded using the *Core Keywording Strategy: Data Collection for a Register of Educational Research* 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 keywords and the review-specific keywords are given in Appendix 2.4.

For those studies where it was possible to obtain full texts, keywording 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. A sample of studies (20%) were keyworded by EPPI-Centre staff for quality assurance. Once differences were reconciled and reasons for exclusion were recorded for each study.

Keywording resulted in the exclusion of a number of studies for not meeting the inclusion criteria, the reasons being recorded and indicated in Chapter 3. The agreed keywords for the remaining studies were used to produce the systematic map of included studies.

2.2.5 Quality assurance process

Records were made of all searches: electronic database searches were documented

and 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 studies (15%, comprising 65 of 431 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 (10 of 48 studies).

2.3 In-depth review

2.3.1 Moving from broad characterisation 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 32 keyworded studies were equally relevant to the review questions and should be included in the in-depth data-extraction. Two studies were later excluded during the data-extraction stage, for different reasons, which were not related to additional principles for exclusion. (During the in-depth data-extraction, it became evident that one study overlapped to too great an extent with another by the same author and, in the second case, a border-line decision for inclusion was tipped toward exclusion by in-depth review.) A revised map of the 30 remaining studies was then created.

2.3.2 Methods for extracting data from studies in the in-depth review

The 30 keyworded studies were entered into the EPPI-Centre's detailed data-extraction software, EPPI-Reviewer, using the *Review Guidelines for Extracting Data* (EPPI-Centre, 2002b). Review-specific questions relating to the weight of evidence of each study in the context of the review were used in addition to those of the EPPI-Reviewer.

Data were extracted from all studies by at least two people working independently. The author extracted data from all 30, while, for the second data-extraction, studies were shared among eight others, including members of the Review Group, members of EPPI-Centre staff and a research assistant.

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

In order to ensure that conclusions were based on the most sound and relevant evidence, judgements were made using the EPPI 'weight of evidence' criteria. This involved judgements about three aspects of each study (A, B and 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 were 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 its soundness. The judgement of methodological soundness was thus dependent on what was reported in the study. The lack of information about a certain feature did not necessarily mean that this feature was not attended to in practice by the study, just that it was not reported by the author of 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.

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 research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?* The concern with reliability and validity of the assessment presented the most straightforward organisation for bringing together the data from the in-depth analysis of the 30 studies. The main problem was that, in practice, a clear-cut distinction between these variables is not easy to establish and some studies explicitly dealt with both. Not only

are reliability and validity interdependent qualities of assessment procedures and instruments, but there are issues surrounding the definition of each, as discussed in section 1.2. It was, however, found possible to designate each one as providing evidence *primarily* in relation to reliability or *primarily* in relation to validity, while recognising that, in many cases, those reporting on reliability also provided, implicitly or explicitly, information about validity, with reliability being a prerequisite for validity. Where studies give evidence for both reliability and validity, they are discussed in detail under 'reliability', with further references in the section on validity.

Thus the two main section of the synthesis were concerned with studies giving evidence manly of reliability or mainly of validity. Evidence relating to the subsidiary question (*What conditions affect the reliability and validity of teachers' summative assessment?*) was discussed as a third section of the synthesis.

2.3.5 In-depth review: quality assurance procedures

All in-depth data-extraction was carried out independently by at least two people, using the EPPI-Reviewer (*Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research* (EPPI-Centre, 2002b)) and the review-specific questions. The author extracted data from all 30 studies in the in-depth review. For ten studies, these data-extractions were moderated by EPPI-Centre staff. Differences were reconciled by telephone. For the remaining 20 studies, the second data-extraction was carried out by a member of the Review Group (five of whom took part) or a research assistant. Again, telephone conversations were used to talk through and reconcile differences by reference to the evidence in the studies.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

This chapter presents results of the stages of searching and screening, using inclusion and exclusion criteria, and the application of the EPPI-Centre and review-specific keywords. The numbers of studies at the various stages of the progression filtering of studies are given in a flow diagram of the process. The characterisation of the selected studies in terms of the keywords is described and the results are given of the quality assurance procedures for this part 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 431.

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

Table 3.1: Results of initial search (431 articles)

Identification	Number (%)	Number included (%)
ERIC	311 (72)	5 (17)
BEI	32 (7)	0
Electronic Database (ERSDAT)	1 (0)	0
Handsearch (not JOL)	49 (11)	17 (57)
Journal on Line (JOL)	4 (1)	1 (3)
Contact	12 (3)	2 (6)
Citation	22 (5)	5 (17)
Totals	***431***	***30***

The criteria for excluding papers and the number excluded at all stages are given in Table 3.2. Three hundred and sixty-nine papers were excluded, some being excluded for more than one reason, while 14 others were unobtainable.

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

	Criteria (more than 1 can apply)	Numbers
Criterion A	not reliability or validity	125
Criterion B	not summative assessment (aptitude and special needs assessment tests excluded)	54
Criterion C	not teacher assessment (assessment of teachers and school evaluation excluded)	275
Criterion D	not school (higher education, nursing education, other vocational excluded)	42
Criterion E	not research (instrument development excluded; also handbooks and reviews)	122

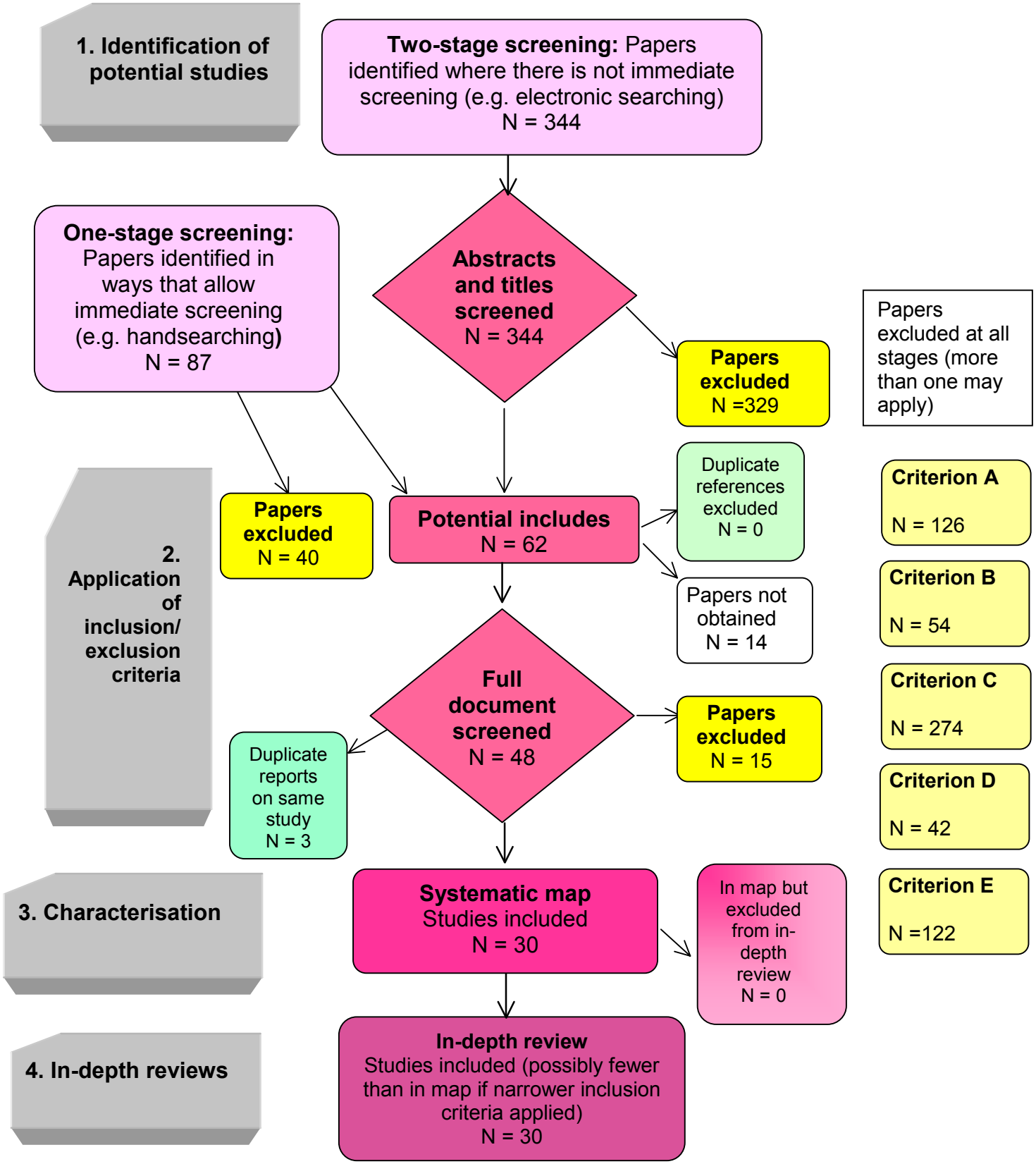
In the screening process all studies were labelled either IN or OUT with the reasons for exclusion. In addition, some studies, considered to be of particular relevance but excluded for one of these reasons, were labelled as USEFUL for the background discussion. Of the 62 studies labelled IN, the full texts of 14 could not be found, leaving 48 for the keywording stage. At this stage (and, in two cases, at the in-depth review stage) 15 further studies were excluded, using the same criteria as above. In addition, it was judged that, in two sets of studies, the same data were used; in one case, three studies were linked and only one included in the data-extraction; in the other, two studies were linked, leaving one for data-extraction. Thus 30 studies remained for in-depth review.

3.2 Characteristics of the included studies (systematic map)

3.2.1 Characterisation in terms of the EPPI-Centre keywords

The classification of the 30 included studies in terms of the keywords is given in Appendix 3.1. Tables A3.1.1 to A3.1.4 give the classification according to the EPPI-Centre keywords. These show that half of the studies were conducted in England, 12 in the United States and one each in Australia, Israel and Greece. 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. Eighteen studies were classified as ‘exploration of relationships’, 11 as ‘evaluations’ and one as ‘description’ of the process of using an assessment, not the assessment as a method, which would have been excluded.

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



3.2.2 Characterisation in terms of the review-specific keywords

Tables A3.1.5 to A3.1.10 in Appendix 3.1 refer to the review-specific keywords. All but a few studies dealt with some aspect of English, mathematics or science. Eighteen 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 nine were either set externally or set by the teacher to external criteria. The majority were assessed by teachers using external criteria. The most common purpose of the assessment in the studies was for national or state-wide assessment programmes, with six related to certification and another six to informing parents. In some cases, there were several purposes and Table A3.1.11 shows that the most common combinations were (i) for informing parents, to be combined with other purposes, and (ii) for monitoring to be combined with national and state testing.

Table A3.1.12 shows the number of studies in which there were various combinations of teacher assessed tasks and type of scoring. Table A3.1.13 and A3.1.14 show how aspects assessed and types of study varied with the educational setting of the studies. There was no variation across educational setting in relation to the focus of the study on reliability or validity, but there were slightly more evaluations of naturally-occurring situations in primary schools. This presumably reflected the interest in assessment of younger children, which was introduced in the 1990s. Almost all studies in the primary and nursery school involved assessment of mathematics and a high proportion related to reading. At the secondary level, studies of assessment of mathematics and 'other' subjects (variously concerned with foreign languages, history, geography, Latin and bible studies) predominated.

Tables A3.1.15 – 17 show how the areas of achievement assessed, the type of tasks and the types of scoring varied with the purpose of the assessment. These tables highlight the predominance of the research relating to national assessment and the use of external criteria by teachers. As might be expected in the context of summative assessment, there is little research on student self-assessment and on teachers using their own criteria.

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 65 of the 431 studies (15%). There was disagreement in five cases, which led to clarification of criterion E.

3.3.2 Keywording

For keywording, where all studies were classified by two people, complete agreement was found for 48 of the 62 studies. The main difference was in relation to the type of study. It appeared that it was possible to apply the category of

'researcher-manipulated evaluation' and 'exploration of relationships' to some studies, depending on whether the process of setting up the study or the outcome of the study was seen as the focus of the classification. Clarification was reached through giving priority to the purpose of the study i.e. in some cases, although there was researcher manipulation, this was for the purpose of exploring the relationship between variables and so was classified as exploration of relationships. A further difference in relation to 'description' was clarified by distinguishing between description of an assessment procedures (that is, what it involved) and description of how it is implemented in practice. The latter was included and the former excluded.

4. IN-DEPTH REVIEW: RESULTS

This chapter describes the characteristics and findings of the finally selected studies. The synthesis of findings in relation to the main review question is given in two main sections dealing with the studies whose main focus is the reliability or the validity of the assessment procedures described. Each of these sections is sub-divided according to the type of evidence which is used. There is a summary of main points at the end of each section. Findings addressing the subsidiary question are brought together under eight sections, 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 conclusion of the 30 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 involved 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 main focus, age of students and weight of evidence

Item	Reliability/ validity reported	Age of learners (years)	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Abbott *et al.* (1994) Some sink, some float: National Curriculum assessment and accountability	Reliability	5-10	Medium	Medium	Medium	Medium
Bennett *et al.* (1993) Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill	Validity	5-10	High	High	High	High
Brown *et al.* (1996) The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations	Validity	17-20	Medium	Medium	Medium	Medium
Brown *et al.* (1998) An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England	Validity	17-20	High	High	High	High
Chen and Ehrenberg (1993) Test scores, homework, aspirations and teachers' grades	Validity	11-16	Low	Medium	Low	Low
Coladarci (1986) Accuracy of teacher judgements of student responses to standardised test items	Validity	5-10	High	High	High	High
Crawford *et al.* (2001) Using oral reading rate to predict student performance on statewide achievement tests	Validity	5-10	Medium	Medium	Medium	Medium
Delap (1994) An investigation into the accuracy of A-level predicted grades	Validity	17-20	High	Medium	Medium	Medium
Delap (1995) Teachers' estimates of candidates' performances in public examinations	Validity	17-20	Medium	Medium	Medium	Medium
Frederiksen and White (2004) An application of validity theory to assessing scientific inquiry: making formative assessment the foundation of school accountability	Reliability	11-16	Medium	High	Medium	Medium
Gipps *et al.* (1996) Models of teacher assessment among primary teachers in England	Validity	5-10 11-16	Medium	Medium	High	Medium
Good (1988a) Differences in marks awarded as a result of moderation: some findings from a teacher-assessed oral examination in French	Reliability	11-16	High	High	High	High
Good and Cresswell (1988) Can teachers enter candidates appropriately for examinations involving differentiated papers?	Validity	11-16	Medium	High	Medium	Medium

Item	Reliability/ validity reported	Age of learners (years)	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Hall and Harding (2002) Level descriptions and teacher assessment in England: towards a community of assessment practice	Validity	5-10	Medium	Medium	Medium	Medium
Hall *et al.* (1997) A study of teacher assessment at Key Stage 1	Validity	5-10	Medium	High	High	High
Hargreaves *et al.* (1996) Teachers' assessments of primary children's classroom work in the creative arts	Reliability	5-10 11-16	High	High	High	High
Hopkins *et al.* (1985) The concurrent validity of standardised achievement tests by content area using teachers' ratings as criteria	Validity	5-10 11-16	High	High	Medium	Medium
Koretz *et al.* (1994) The Vermont Portfolio Assessment Program: findings and implications	Reliability	5-10 11-16	High	Medium	High	High
Levine *et al.* (1987) The accuracy of teacher judgement of the oral proficiency of high school foreign language students	Reliability	11-16	High	High	Medium	Medium
Meisels *et al.* (2001) Trusting teachers' judgements: a validity study of curriculum-embedded performance assessment in kindergarten to grade 3	Validity	0-4 5-10	High	High	High	High
Papas and Psacharopoulos (1993) Student selection for higher education: the relationship between internal and external marks	Validity	17-20	Medium	Medium	Low	Low
Radnor (1995) Evaluation of Key Stage 3 assessment in 1995 and 1996. Evaluation of Key Stage 3 assessment arrangements for 1995	Validity	11-16	Medium	Medium	Medium	Medium
Reeves *et al.* (2001) The relationship between teachers assessments and pupil attainments in standards test tasks at Key Stage 2, 1996-1998	Reliability	11-16	High	High	High	High
Rowe and Hill (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles'	Reliability	0-4 5-10 11-16	Medium	High	High	High
Shapley and Bush (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience	Reliability	0-4 5-10	High	High	High	High

Item	Reliability/ validity reported	Age of learners (years)	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Sharpley and Edgar (1986) Teachers' ratings vs standardised tests: an empirical investigation of agreement between two indices of achievement	Validity	5-10	High	High	Medium	Medium
Shavelson *et al.* (1992) Performance assessments: political rhetoric and measurement reality	Reliability	5-10 11-16	High	Medium	Low	Low
Shorrocks *et al.* (For NUT and University of Leeds) (1993) Testing and assessing 6 and seven year-olds. Evaluation of the 1992 Key Stage 1 National Curriculum Assessment Final Report	Reliability	5-10	High	High	High	High
Thomas *et al.* (1998) Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment	Reliability	5-10	High	High	High	High
Wilson and Wright (1993) The predictive validity of student self-evaluations, teachers; assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary schools	Validity	11-16 17-20	High	High	Medium	Medium

4.2 Synthesis of evidence: overall review question

What is the evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?

The most straightforward way of bringing together the data from the 30 selected studies to answer the main review question is to consider evidence for reliability and validity separately. In practice, a clear-cut distinction between these variables is not easy to establish. Not only are reliability and validity interdependent qualities of assessment procedures and instruments, but there are issues surrounding the definition of each, as discussed earlier. However, as a basis for identifying patterns in the evidence from the studies, each study has been designated as providing evidence primarily in relation to reliability or primarily in relation to validity. This is done while recognising that, in many cases, there are those reporting on reliability who also, implicitly or explicitly, provide information about validity, since reliability is seen as a prerequisite for validity. Where studies give evidence for both reliability and validity, they are discussed in detail under 'reliability', with further references in the section on validity.

To be explicit about the operational basis for classification, studies are labelled as mainly concerned with reliability where there is re-scoring or re-marking or moderation of the data from the assessment process, or where the purpose is exploration of the influence of school or student variables on the assessment outcome. Studies are labelled as mainly concerned with validity where they report the relationship between one score and another intended to measure the same achievement, or where there is a study of variation in procedures and in information used in the assessment which is relevant to the question of what was being assessed. Using these boundaries, the classification of the studies is given in Table 4.1.

4.2.1 Evidence from studies concerned with the reliability of teachers' assessment

As Table 4.1 shows, of the 12 studies concerned mainly with reliability, there were eight providing evidence of high overall weight, three providing evidence of medium weight and one giving low weight evidence in relation to this review. Within these studies, there were some where the evidence for teachers' judgements came from regular classroom work in particular subject areas, and others where evidence came from specified tasks as in the case of science investigations or an interview in a foreign language.

Studies using evidence from regular classroom work

Six studies providing evidence of high weight and one giving evidence of medium weight are included in this section. Information about the age group and areas of achievement assessed, study type and overall weight of evidence is brought together in Table 4.2. The studies provide different kinds of evidence of reliability. The evidence in Koretz *et al.* (1994), Shapley and Bush (1999), and Rowe and Hill (1996) is derived from rescoring the work assessed by teachers. Reeves *et al.* (2001), Thomas *et al.* (1998), and Rowe and Hill provide evidence relating to the consistency

with which assessment criteria are applied by teachers, while Abbott *et al.* (1994) report on the consistency of administration of performance assessment.

Table 4.2: Studies providing information of reliability of assessment by teachers based on regular classroom work

Study	Age of learners (years)	Achievement assessed	Study type	Overall evidence weight
Koretz *et al.* (1994) The Vermont Portfolio Assessment Program: findings and implications	5 - 16	Reading Maths	Evaluation: naturally-occurring	High
Shapley and Bush (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience	3 - 8	Reading Writing Maths	Evaluation: naturally-occurring	High
Rowe and Hill (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles'	3 - 16	Reading Writing Maths	Evaluation: naturally-occurring	High
Reeves *et al.* (2001) The relationship between teachers assessments and pupil attainments in standards test tasks at Key Stage 2, 1996 - 1998	11 - 12	Reading Writing Maths Science	Exploration of relationships	High
Thomas *et al.* (1998) Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment	6 - 7	Reading Writing Maths Science	Exploration of relationships	High
Shorrocks *et al.* (1993) Testing and assessing 6 and seven year-olds. Evaluation of the 1992 Key Stage 1 National Curriculum Assessment Final Report	6 - 7	Reading Writing Maths Science	Evaluation: naturally-occurring	High
Abbott *et al.* (1994) Some sink, some float: National Curriculum assessment and accountability	6 - 7	Science	Evaluation: naturally-occurring	Medium

Two high-weight studies concerned assessment systems which involve teachers assessing portfolios of students' work. Koretz *et al.* (1994) is one of several published accounts of the Vermont portfolio system. Three other papers reporting this work (Koretz 1998, Klein *et al.*, 1995 and Koretz *et al.*, 1991) were found in the review search. Data-extraction was carried out only for Koretz *et al.*, 1994.

Vermont's assessment programme was created in the late 1980s to provide 'high-quality data about student achievement (in this case, sufficient to permit comparisons of schools or districts) and to induce improvement of instruction' (Koretz *et al.*, 1994, p 5), while at the same time avoiding the negative consequences of test-based accountability systems. At the time of the study by Koretz *et al.* (1994), the

programme was limited to Grades 4 and 8 in writing and mathematics. The programme is described as follows:

> 'The centrepiece is portfolios of student work that are collected over the course of a year by classroom teachers. Teachers and students have nearly unconstrained choice in selecting tasks to be placed in the portfolios. The program is truly "bottom up" and committee consisting mostly of teachers have had primary responsibility for developing the operational plans for the program, constraining teachers' and students' choices and designing scoring rubrics. The portfolios are complemented by "uniform tests" which are standardised but need not be multiple choice' (Koretz *et al.*, 1994, p 6).

For the assessment of mathematics, teacher and students selected five to seven 'best pieces' from the portfolio for scoring. This scoring was done on a four-point scale for each of several aspects of the pieces of work. The scoring for the mathematics work was carried out by teachers other than the students' teachers, in regional meetings in 1991–1992 and in a single state-wide meeting in 1992-1993. In the first year, teachers scored their own students' work in writing, but, in the second year, this was done by other teachers.

In their study, Koretz *et al.* (1994) selected a random sample of portfolios for re-scoring. Over the two years, 1991–1993, the number of portfolios double-scored ranged from 161 for eighth-grade mathematics and 779 for fourth-grade writing. The varying number was partly due to the greater time required for scoring mathematics and to a number of portfolios being incomplete and removed from the sample.

The main findings of relevance to this review relate to the rater reliability of the scores. However, in a large component of the study, the researchers interviewed staff in a random sample of 80 schools. The results of this were encouraging in relation to the programme's goal of improving instruction. On the other hand, as the authors report:

> 'The positive news about the reported effects of the assessment program contrasted sharply with the empirical findings about the quality of the performance data it yielded. The unreliability of scoring alone was sufficient to preclude most of the intended uses of the scores' (Koretz *et al.*, 1994, p 7).

The results were that the rater reliability was very low in both writing and mathematics in the first year of the study. It improved appreciably for mathematics in the second year (1993) but not in writing. The authors based their conclusion on correlations rather than percentage agreements between ratings, pointing out that the latter can lead to errors when a scale has few values (as these did). Although the Vermont programme was not intended to provide student-level scores for external use, the study investigated these since 'the reliability of student-level scores places a bound on the quality and validity of all of the assessment results, whether individual or aggregate' (Koretz *et al.*, 1994, p 7).

Results were analysed at three levels: (i) the score for each piece in the portfolio on each of the scoring dimensions (7 in mathematics and 5 in writing), (ii) the dimension level, combining scores across pieces, and (iii) the portfolio level. For writing, all correlations between raters were similar at these three levels and all hovered around 0.40 for both years. In mathematics, piece-level correlations were low in the first year

and improved in the second. Correlations at the dimension level were higher, reaching 0.79 for the eighth-grade students. Generalisability showed that much of the variance in scores in both writing and mathematics could be attributed to disagreements among raters. In other respects, they were different. In mathematics, unlike in writing, the performance varied from piece to piece. This meant that 'a larger number of pieces will be needed to obtain reliable score in mathematics' (Koretz *et al.*, 1994, p 9).

Koretz *et al.* (1994) concluded that two factors contributed to the unreliability of the scores: (i) the difficulty of training a large number of raters and (ii) the lack of standardisation of tasks. This lack of standardisation required raters to stretch general-purpose rubrics to cover a wide variety of tasks. They point out that many performance-assessment programmes that have demonstrated high levels of rater reliability have relied on standardised tasks and have used task-specific rubrics. An intermediate approach would be to allow unstandardised tasks but to apply genre-specific scoring rubrics.

There is a considerable degree of consistency between the findings of Koretz *et al.* (1994) and those of Shapley and Bush (1999) in their study of the Dallas Public Schools' reading/language arts portfolio assessment for pre-kindergarten to Grade 2. As in Vermont, portfolio assessment was introduced in the early 1990s and, at first, was 'informal and unstructured and could best be described as "a classroom collection of students' work'. The guidelines simply asked teachers to include samples of students work and a grade-level checklist in a folder' (Shapley and Bush, 1999, p 114). Later, however, in 1994, procedures were tightened and 'instructional specialists and evaluation staff' drew up guidelines for selecting work, scoring and recording. The 1995-96 version of the portfolio – the subject of Shapley and Bush's (1999) study – required the portfolio components to be aligned with the state content standards (curriculum goals and criteria) and to use the same scoring criteria for all students at a particular grade, regardless of special needs status.

The guidelines specified that at least 12 samples of work relating to the state curriculum and covering a variety of types of work were to be collected by teachers over the year. The teachers scored their own students' portfolios, according to a four-point scale for each curriculum goal. In addition, they made 'an overall judgement of how well a collection of work samples meets the multi-dimensional standards that define each instructional goal' (Shapley and Bush, 1999, p 117). A random sample of portfolios was re-scored by other teachers, who were randomly selected from those who had attended the district portfolio training and had themselves implemented portfolios that year. The re-scorers were given extra training and were paid for the work. During rating, the second raters made notes about the adequacy of each portfolio and whether, for instance, teachers had documented the instructional goals as required.

Ratings given to students' work by their own teachers were higher than those given by the second raters, in all cases. The difference reached significance for kindergarten (Kg) and pre-Kg for all instructional goals except for 'writing about experiences'. For grades 1 and 2, all the differences were significant except for listening and speaking. In terms of consistency, the percentage agreement between teachers and second raters was between 50-59% for grades 1 and 2 portfolios, varying with the goal, and when raters were not in perfect agreement, a difference of

one point was most likely. However, as Koretz *et al.* (1994) pointed out, with only four points, this is not a good indicator of agreement. Inter-rater correlations were in fact low. A mean pre-Kg and Kg correlation of 0.37 indicated that 14% of the variance in second raters' scores was explained. Likewise, the 1st and 2nd grade mean correlation of 0.48 indicated that 23% of the variance in second raters' scores could be predicted from knowing the teachers' scores. Large percentages of unexplained variance were due to error.

As found by Koretz *et al.* (1994), the low reliability of scoring made judgements of validity problematic. Moreover, ratings could not be made about half of the time for many instructional goals because of inadequate evidence. 'Overall, for many portfolios, there was insufficient sampling of the content domain because the number of samples was inadequate, the work sample provided inadequate information, the purpose for the work samples was unknown, the work samples did not exemplify the goals' content knowledge, or there were no teacher notations to explain activities and to clarify the student's performance' (Shapley and Bush, 1999, p123).

Thus Shapley and Bush concluded that, after three years of development, the portfolio assessment did not provide high quality information about student achievements for either instructional or informational purposes. They suggested that the unreliability of the scores was likely to be related to (a) lack of standardisation of tasks, (b) problems with the scoring rubrics and (c) inadequate training. Even though the training and guidelines were more stringent than in the case of the Vermont portfolio programme, these may have contributed to the low reliability. However Shapley and Bush (1999) emphasise the role of the scoring rubrics in relation to the tasks.

> 'Because the tasks were unstandardised, the scoring rubrics aligned with the instructional goals rather than with specific tasks....Improving the scoring rubrics will require greater standardisation of the portfolio contents so that there is stronger alignment between the tasks and specific evaluative criteria. This suggests a need for a compromise between standardisation, which is needed to improve technical quality, and the flexibility that allow portfolios to be integrated with the classroom context' (Shapley and Bush, 1999, p 127).

Further, the authors suggest that it seems that portfolios need to contain a core of essential work samples (those that all portfolios must contain) and optional work samples (those that the teacher and students agree to select). The core samples would provide a common frame of reference across all portfolios for judging students' performance.

The approach to supporting TA reported by Rowe and Hill (1996) differs from the suggestions made by Shapley and Bush (1999). The 'subject profiles' used in Victoria, Australia, since 1986, rather than closely defining the tasks to be assessed, provide a more detailed description of the criteria to be applied in relation to regular classroom work. The term 'subject profile' refers to a framework for helping teachers, schools and systems in assessing and recording students' educational progress. The framework comprises a set of indicators of competency which have been 'empirically validated and calibrated on a common scale, thus enabling use of the full range of assessment methods available to teachers' (Rowe and Hill, 1996, p 318). Such indicators are provided for each aspect of each curriculum area. Each set of

indicators is arranged in a sequence of developing competency, following the lines of ‘progress maps’ for developmental assessment (Masters and Forster, 1995). The profiles were designed ‘to provide a means whereby teacher assessments of student performance on the curriculum as taught in schools could be reported using criteria that are consistent across classrooms, schools and the system’ (Rowe and Hill, 1996, p 326).

The empirical studies reported by Rowe and Hill were part of research and development projects carried out between 1988 and 1994. In the project from which the data were reported, the Victorian Quality Schools Project, 1992–1994, teachers were requested to rate their students' levels of achievements with reference to the indicators for each of the nine levels or bands of the reading, writing and spoken language strands of the English profiles, and for each of the twelve levels (1 to 12) of the number and space strands of the mathematics profiles. The results were recorded on class recording lists, where a score of 3 was typically recorded if all the behaviours associated with a given band/level were consistently displayed by the student; 2, if most of the behaviours were present; 1 if some of the behaviours are beginning to be developed; and 0 if none of the behaviours have yet been observed. The ratings for each band/level were then added together to give a total score out of 27 for each English profile strand, or 36 for either strand of the mathematics profiles (Rowe and Hill, 1996, p 327). Entire year cohorts from 90 schools, Kg to Grade 12, were involved.

Since the levels rest on the assumption of a cumulative scale, the Guttman alpha coefficients were used as reliability estimates, calculated separately for each year group. In addition, for the reading strand, retest data (a repeat of the rating by the same teachers) were reported and the results of second raters were obtained in situations where teachers co-taught.

The Guttman reliability estimates, with coefficients ranging from 0.77 to 0.96, indicated that the profiles appeared to function as cumulative scales or growth continua and that teachers were consistent in their use of them. 'The reliability coefficients were not as high for early years as for later years, on account on the restricted range in the achievement levels of students in earlier years. In addition, with the exception of the preparatory grade (Kg), estimates for the two mathematics strands are somewhat higher than for the three English strands' (Rowe and Hill, 1996, p 128). Pearson product-moment correlations between teacher assessments of the same students made on two occasions four months apart, indicated high test-retest reliability (correlations values from 0.89 to 0.93). Inter-rater reliability (different teachers, same students) were also high (0.83 to 0.89), based on an opportunity sample.

The authors concluded that, when using subject profiles, teachers are consistent in their assessment of students and are also able to achieve a satisfactory level of inter-rater reliability, although the evidence was only partial on this point. They were confident that ‘profiles allow teachers to communicate to parents about student progress and achievement using a language and standards which are consistent across classrooms, schools and school systems’ (Rowe and Hill, 1996, p 335).

Three studies addressed the reliability of teachers’ judgements in the context of the National Curriculum Assessment (NCA) in England and Wales through comparison

of teachers' assessment (TA) with results of standard tests and tasks. In these studies, there were no visits to classrooms, data being derived from students' scores and from questionnaires to teachers. Thus there was no information about how teachers made their judgements, in contrast with Abbott *et al.* (1994) who report observations made during test administration, and with the studies of Hall and Harding (2002), Hall *et al.* (1997) and Gipps *et al.* (1996), which are discussed later in the section on validity. [In all the studies of the NCA, the date of data collection has to be borne in mind, since the form of the tests and the requirement for assessment by teachers (TA) for 6 and seven year-old pupils (Key Stage 1) changed during the period 1991 to 1994. In 1991 teachers had to assess their pupils against a number of criteria (statements of attainment) and also to administer standard performance tasks which assessed a number of core attainment targets. The latter were extremely time consuming and in 1992 were replaced by much shorter performance tasks and some paper and pencil tests. The burden of TA was also reduced after a review of NCA in 1993 and 1994 which resulted in the introduction in 1995 of 'level descriptions' to replace the statements of attainment. For each attainment target, teachers were to judge achievement against level descriptions, using a 'best fit' approach.]

Reeves *et al.* (2001) reported data collected by the School Sampling Project, a longitudinal project started in 1995 which 'tracks pupil performance and schools' implementation of the curriculum over a period of time based on a sample framework of 1000 schools' (Reeves *et al.*, 2001, p 143). In the NCA, students are assigned levels of 1 to 8, 4 being the target level for 11 year-olds. Levels based on the standard test scores for 11 year-olds in reading, writing, mathematics and science were collected. A TA level for each student was also recorded, derived from levels assigned for each attainment target according to a specific formula. It is intended that the TA should cover the full range of the curriculum for English, mathematics and science and include work in a variety of contexts. Thus evidence from activities in mathematics and in science involving investigation over a period of time can be included as can pupils' use of language in debates or role-playing activities. Since such a wide range of contexts cannot be encompassed in the tests, it is not expected that there will be complete agreement between the level based on the test and TA. Discussing this point, Reeves *et al.* (2001, p 142) point out the following:

> 'if the methods agree completely in nearly every case, there would be a strong argument that one or the other was redundant, while, on the other hand, if they frequently yield quite different results, this would raise serious concerns that the system contained a fundamental flaw.'

Using data collected in 1996, 1997 and 1998, Reeves *et al.*, (2001, p 153) reported that comparisons between test results and TA 'reveal a remarkably high level of consistency across years in all three subjects' (Reeves *et al.*, 2001 p 153). But the direction of the differences varied across subjects. In mathematics, test levels were lower than TA (the difference being significant for 1996 and 1998 but not 1997), while in English and science the opposite was found. However, although significant, the differences by subject were small. The proportion of exact agreement between TA and test results remained consistently around three-quarters (75%) and less than 0.5% of disagreements exceeded one level.

Analysis of variance was used to explore the relationship between school and student characteristics and the size of the difference between TA and test scores. This revealed that the school had a big impact, but the amount of variance explained by this factor declined over time in all subjects, most notably science. In relation to student gender, the difference was significant for mathematics across all years, with the TA consistently under-rating boys more than girls. The same was found for science; but, for English, the opposite was found, where females were more frequently under-rated by the TA. For students with special educational needs (SEN), TA levels were more likely to be lower than test results. In many instances the effect was considerable, for example, 24% of students with SEN were awarded lower levels than the test results for science in 1998. For students whose first language was not English, the only effects were in English in 1998, when TA under-rated 25% of these students compared with 15% of others.

The authors conclude from the consistency across the years and the size of the extent of agreement between test results and TA that it 'would seem to fit the bill of having two complementary assessment measures which usually concur, but which vary sufficiently to justify maintaining the use of both' (Reeves *et al.*, 2001, p 158). Subject, gender, age and English as a second language had some varying effects on the difference between TA and test scores. However, the strongest relationship was in relation to students with SEN, where tests frequently exceeded their TA levels. This happened in all subjects, but particularly in English and science. The authors offer an alternative explanation to the obvious one for this - that teachers 'teach to the test' for these students, who therefore may be able to perform in the tests above their real attainment level.

Thomas *et al.* (1998) also explored the relationship between TA and performance on standard tests/tasks within a study that was primarily concerned with the relative performance of groups of students differing by gender, linguistic and socio-economic background and special needs. This study used data from the NCA of seven year-olds in 1992 in England and Wales. At that time, standard tasks were administered to individual students and scored by teachers following specified rules. For nine areas (four for English, three for mathematics and two for science), performance was assessed by both TA and standard tasks. Data were also available for the same sample of students for a standardised word-recognition test (WRT). Multilevel modelling was used as the main method of data analysis for exploring the relationship between background characteristics and the three measures.

The results of exploring the relationship between student characteristics and TA and standard tasks separately were that, for 13 out of 63 comparisons, there were greater differences between groups for TA; while for the standard tasks, differences were only greater for three comparisons. Although, overall, TA and standard tasks 'worked similarly', TAs were more likely to widen the gap between groups of students, particularly between those with and without a statement of special needs.

The comparison between TA, standard task and WRT results showed a fairly strong positive relationship overall, with correlations ranging from 0.92 for reading and 0.77 for the 'mathematics probabilities' scores. However, further multilevel modelling, designed to establish the impact on TA of student background factors and standard scores, suggested that 'across all Attainment Targets, teachers are systematically assessing students differently on the TA in comparison to the standard task and that

schools (and teacher judgements) vary in the way TA are scored, even after standard task results are taken into account' (Thomas *et al.*, 1998, p 230). The results show that students' Standard Task assessments account for between 59% and 94% of the variation between schools in the teacher assessments. But there was unexplained school level variation which suggest that 'certain aspects of how teachers judge student outcomes… need to be examined in more detail. Indeed, when student background factors are added to standard task attainment in the model predicting TA, findings suggest the possibility of systematic teacher bias.' (ibid). In most cases, once the standard task levels have been accounted for, each student background characteristic still has a small but statistically significant impact on TA level after their standard task attainment has been taken into account.

The purpose of the study conducted by Shorrocks *et al.* (1993) for the National Union of Teachers and University of Leeds (1993) was to evaluate the results of the tests and the views of teachers on the conduct of the 1992 national assessment for seven year-olds (the assessments also studied by Thomas *et al.*, 1998). The same team had evaluated the 1991 tests for seven year-olds and were able to make comparisons between the years. This review extracts only the data relating to the comparison of scores for the TA and standard tasks and the variation with student characteristics. A representative sample of schools was involved, with 128 teachers completing questionnaires and performance data collected for 1,766 students.

The analysis was in terms of the observed agreement between TA and standard task results. The authors caution that a proportion of individual children may be placed in different levels by the two assessments, even though the overall numbers at each level may be the same. There was close agreement for the four English Attainment Targets but less for the mathematics and science attainment targets. The authors note that the levels of agreement were considerably greater than those they found in 1991. Unlike Reeves *et al.* (2001), the authors do not discuss an optimum level of agreement, apparently assuming that complete agreement is desirable.

In relation to student characteristics, Shorrocks *et al.* (1993) found that, at the subject level, the TA and standard task results behaved in the same way. For gender of student, there were statistically significant differences in favour of girls in English and no differences at subject level in mathematics and science. Older students had statistically higher scores at the subject level for English, mathematics and science, as did those from higher neighbourhood status areas. The ethnic group of students was associated with significant differences only in English, where white children appeared to be superior. For language background at the subject level, there were significant differences in mathematics and science in favour of those with English as a first language. Differences in favour of English-speaking children were found in some aspects of all subjects. In all subjects, there were significant differences, in favour of those without special educational needs (SEN).

The authors note that, in comparison with 1991 results, performance levels were higher. They suggest several reasons for this, such as greater teacher confidence in the assessment procedures and the publication of results in 1992, although not in 1991.This publication raised the stakes and, together with a further year of experience of the national curriculum, could have led to greater encouragement of progress in specific areas or possibly greater teaching to the test.

It is relevant to note at this point the findings of the study by Abbott *et al.* (1994), which concerned the NCA for seven year-olds in 1991, with some discussion of changes in 1992. This study has medium weight, partly on account of its focus on the administration of standard tasks rather than TA, although tasks were administered and marked by teachers. However, it is clearly of relevance to the discussion of the reliability of TA to consider the reliability of the standard tasks, for these are often taken as a benchmark against which variations in TA are compared, as by Thomas *et al.* (1998) and Reeves *et al.* (2001), with the implication that lack of correspondence is due to unreliability of TA.

Abbott *et al.* (1994) observed the administration to seven year-olds of the standard task in science as part of the NCA in England and Wales in 1991. They observed in three classrooms in different circumstances, using open-ended observation methods, field notes and written records of teacher-child and child-child conversations and other interactions. The data they gathered was about the administration of the task (on the floating and sinking of various objects) to all seven year-olds. They found considerable variations among the schools in the conditions of administering the task, in relation to interruptions, help given in the classroom during the administration, etc. While variations in these conditions would be obvious to anyone, the authors note that:

> ‘continuous observation while the SAT took place revealed factors in teachers' presentations of the tasks which made the Government's declared aim of standardising their assessment techniques and children's experience appear completely out of reach’ (Abbott *et al.*, 1994, p 163).

At the same time, although they considered that the task was 'extremely unreliable' (ibid, p 166), they questioned the decision of the Government to drop the task from the testing programme for the next and subsequent years. They pointed out that the skills assessed were valuable ones which were not addressed in any of the 1992 science tasks. They also note that teachers found other standard tasks equally difficult to use, such as the reading tasks. They found that teachers were worried about the subjective judgements involved in standard tasks procedures. It appeared likely that TA is as trustworthy as standard testing over most areas and has the added advantage of being able to fulfil diagnostic and formative aims. The authors considered that supplementing TA with some form of standardised testing in limited areas in order to increase reliability (for summative purposes) would be likely to mean that teachers concentrated on what is tested. The pointed out that 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' (Abbott *et al.*, 1994, p 171).

Studies using evidence from specific pieces of work

In these studies teachers assessed pupils’ performance in specific activities or types of activity. Two provided evidence of high weight for the review, two medium weight and one low weight. The evidence of inter-rater reliability was given in four of the studies and in the other the evidence related to internal consistency. Table 4.3 sets out the five studies in the sequence in which they are discussed.

Table 4.3: Studies providing evidence of reliability of assessment by teachers of specific pieces of work

Study	Age of learners (years)	Achievement assessed	Study type	Overall evidence weight
Hargreaves *et al.* (1996) Teachers' assessments of primary children's classroom work in the creative arts	5 - 11	Writing Art/music/ dance/PE	Evaluation: researcher-manipulated	High
Frederiksen and White (2004) An application of validity theory to assessing scientific inquiry: making formative assessment the foundation of school accountability	11 - 14	Science Practical maths/science/ tech	Evaluation: researcher-manipulated	Medium
Good (1988a) Differences in marks awarded as a result of moderation: some findings from a teacher assessed oral examination in French	15 - 16	French (oral)	Exploration of relationships	High
Levine *et al.* (1987) The accuracy of teacher judgement of the oral proficiency of high school foreign Language students	12 - 16	French and Spanish oral proficiency	Exploration of relationships	Medium
Shavelson *et al.* (1992) Performance assessments: political rhetoric and measurement reality	11 - 12	Science	Evaluation: researcher-manipulated	Low

In one of the few studies relating to the assessment of work in the arts, Hargreaves *et al.* (1996) developed and trialled with primary teachers a set of scales derived from the constructs used by teachers in assessing work in visual arts, music and creative writing. The development of the scales used a 'bottom-up' approach, starting from the constructs that teachers use in assessing students' work in the arts. Eleven teachers each carried out with their students a 'structured' and an 'unstructured' activity (selected from a choice of activities) in one of these domains. The outcomes were used in deriving a set of seven-point bipolar scales for each of the constructs. Nine of these teachers then attended a further meeting held later, when they were asked to rate a new set of products in each domain. Some constructs were added at this point, for example, 'aesthetically appealing – unappealing' and 'technically skilful-unskilful' were added to all three lists of constructs, producing 17 constructs for visual arts, 14 for music and 13 for creative writing. Some of these constructs were evaluative, while others were more neutral.

The relationship between teachers' judgements were explored by computing 9x9 product moment correlations matrices between the teachers' ratings of each individual product on each scale separately for each domain. This was done for each of the rating scales for each of the six product categories (visual arts – structured, (VA-S); visual arts unstructured (VA-U), etc.). Mean differences between rating of products from S and U activities were computed. These calculations were also carried out for the evaluative scales only.

The results showed that mean correlations across all scales for the six categories of activities (VA-S, VA-U, M-S, etc.) were all significant at or beyond the 0.05 level. Mean correlations between scales were also significant at the 0.01 level or beyond.

The mean correlations for structured activities in visual arts and music were greater than for the unstructured activities, but the reverse was the case in writing activities. In most cases, the ratings of unstructured activities were higher than for structured activities. When a t-test of this difference was repeated using only the evaluative scales, the difference remained, with products of U activities being rated more highly than S activities, and several differences reaching significance level.

The authors conclude from this relatively small-scale project, that teachers can use the 'vocabulary of assessment' in a consistent fashion. The higher level of agreement for the 'structured' than for the 'unstructured' activities in the case of the visual arts and music domains can be explained by the former giving rise to a more uniform set of products. This was not the case for the writing activities, for which the level of agreement was higher for the 'unstructured' activities. The high level of inter-correlations between scales across teachers suggests that teachers were applying all the scales in essentially the same way. They claim that the study 'has demonstrated that when teachers are given the opportunity to clarify their ideas and the ambiguities of language used to describe children's work, they are capable of substantial agreement about the quality of different pieces of work from different students, and apparently make these assessments in uni-dimensional evaluative terms' (Hargreaves *et al.*, 1996, p 210). They also note that 'the more explicitly teachers define the end-product of the activity which they set, the more rigorous they seem to be in assessing the quality of this work' (ibid).

A somewhat similar approach, using teachers to participate in producing a scoring scheme and then involving them in using it, was used by Frederiksen and White (2004) in relation to assessing students' science projects. Although providing evidence of medium weight on account of lack of some detail of the research, this study is of particular significance for this review since it attempts to show how classroom assessment 'primarily intended to promote learning, could become an important source of information for evaluating a school's effectiveness within an accountability system' (Frederiksen and White, 2004).

Six middle-school teachers participated in an iterative design process in which they tried out an initial design for the scoring students' science projects program (the computer-based 'Inquiry Scorer') and provided feedback to the developers. Using a revised version of the scoring software, they then scored 16 projects, using each of seven criteria. They scored the projects individually and then met in small groups to discuss their scoring for every fourth project scored. They were also required to develop a 'map' of each project, to identify the design of the project. The measure of agreement among scorers was taken as the percentage of scorers who gave the modal category of response to each question. Correct coding was determined by one of the authors (JF) after reviewing each project and the scorers' responses. The consistency in rating each of the seven criteria for the overall assessments was also computed and teachers' judgements of the overall quality were analysed for consistency.

The average rate of agreement across all of the projects was 81%, with the agreement rates for individual teachers ranging from 76% to 85%. Thus, the teachers were for the most part consistent with each other in coding the features of a project. Teachers' consistency in identifying and naming the independent variable

was lower (73% correct) than for the dependent variable (82% correct) when they developed their project map.

An interesting finding was the comparison between the criterion ratings of the three teachers who were new to scoring inquiry projects with the two teachers who had had prior experience in holistic scoring in an earlier part of the study. The average consistency for the new scorers was 72%, with a range of 63% to 79%. The corresponding average consistency for the teachers with prior experience in holistic scoring was 69%, with a range of 55% to 79%. Thus, using the Inquiry Scorer with its detailed project analysis led to a high degree of consistency among the teachers in rating important dimensions of performance, and the teachers' ability to make such judgements did not depend on their prior experience in scoring such work.

In their conclusions, Frederiksen and White (2004) emphasise that establishing the credibility of judgements of performance depends on the nature of the tasks and not just on inter-scorer reliability. They argue that having open standards for tasks and how they are assessed makes possible a merging of classroom assessment goals and goals for creating evidence of students' learning that can be used by schools in meeting accountability standards. This alignment depends upon having descriptions of types of activities or tasks that provide meaningful challenges to students and teachers to aim for in learning, and also on having transparent processes for evaluating performance that can be used by teachers and students in reflecting on their work.

The studies of Good (1988a) and Levine *et al.* (1987) were of TA in foreign languages. Good's study (and a linked paper, Good (1988b)) concerned the differences in marks given by teachers and moderators in French oral examinations of the GCSE. As part of a study of different ways of grading marks from differentiated examination, teachers were trained in administering and marking oral examinations in French. They administered a trial examination to candidates prior to these students taking the regular examination later in the year. From the candidates involved (177 at 'general' level and 122 at 'extended' level), a random sample of recorded oral examinations was drawn for re-scoring by moderators, who were unaware of the marks awarded by the teachers.

The teachers' marks were generally more generous than the moderators' average mark. At the general level, the mean teachers' mark was 3.1 marks higher than the mean moderators' mark; this was equivalent to 0.4 grades on the oral component of the examination. At the extended level (a more open-ended interview for more advanced students), the mean teachers' mark was 5.3 marks higher than the mean moderators' mark; this was equivalent to 1.1 grades. The correlations indicate that there was no significant difference in variance of the teachers' and moderators' marks and the two agreed on the rank order of candidates. Most of the extended level correlations were lower than the general level correlations.

After considering three different methods of adjusting marks (under different assumptions of no error in the teacher's mark, no error in moderator's mark or that errors occur in both, proportional to their variances), the author suggests that the candidates should be awarded the adjusted teacher's mark rather than the raw teacher's mark or the moderator's mark.

The author concludes that, when given some training in the tasks to be undertaken, teachers conducting French oral examinations are able to place their candidates in a rank order that is consistent with the specified criteria nearly as effectively as assistant examiners marking conventional examination papers. However, since the teachers were more lenient in awarding marks, some adjustment of scores is needed. In practice, whichever version of the general statistical method is used, there will be appreciable differences only for candidates at the extremes of the achievement range in each centre.

Similar findings emerged from the study by Levine *et al.* (1987) of oral proficiency in French and Spanish of high school students. For a random sample of their students, eight teachers were asked to judge the scores on the American Council on the Teaching of Foreign Languages (ACTFL) oral interview for four of their students, randomly selected. The language proficiency of these students was rated in an oral interview by independent testers who were certified by ACTFL. Personal information about the teachers and about the selected students, including their letter grade (A, B or C) based on class work in the language, was also collected.

It was found that the teachers consistently overestimated their students' ability. There was a significant difference between teachers' predicted ACTFL rating (mean 4.9) and students' actual rating (mean 3.4). The means for the French and Spanish groups were not significantly different. The number of years of instruction in the foreign language did not influence the difference between estimated and actual ratings. In relation to letter grade, there was a definite trend. 'A' students were overestimated by a greater amount than 'B' students who in turn were overestimated more than 'C' students. In relation to groups based on the actual ACTFL score, teachers overestimated more the students performing at the lower end of the scale. The interaction between actual rating and letter grade was approaching significance. The authors suggested that this means that academic letter grade severely biases teacher judgements.

The authors suggest, on the basis of their findings, that teachers need workshop training directly addressed at the consistent overestimation and letter grade influence. They cite evidence that training programmes can make teachers more accurate in their judgements. They claim that 'teacher themselves would be receptive to assessment training…Perhaps the most important by-product of assessment training may be that teachers will begin to modify their curriculum to make day-to-day classroom activities more closely congruent with the increased emphasis on oral language proficiency' (Levine *et al.*, 1987, p 50).

The study by Shavelson *et al.* (1992) is rated as having low weight for this review mainly on account of its main aim being to compare different forms of performance assessment, only one of which involved teachers in observing students and judging their performance. Fifth and sixth grade students in the USA responded to science assessment in the form of observed investigations with notebooks, paper-and-pencil measures on the same topic and computer simulations. Three hands-on investigations ('paper towels', 'Electric mysteries' and 'bugs') were created and treated as the 'benchmark' assessments. When conducting these investigations, the students were observed and they also used notebooks to record specific aspects of the investigations; these were collected as a second mode of assessment. Computer simulations were created for two of the investigations (omitting 'paper towels'). Short-

answer and multiple-choice questions were chosen to parallel in content the three hands-on investigations. The students, 300 in fifth and sixth grade, were selected from two school districts, differing in their science curricula. One district was well known for its 'hands-on' curriculum, while the other had no regular science, apart from what was taught as part of a text-book course on health (Shavelson *et al.*, 1992, p 23).

It is implied that the benchmark investigations were observed by two raters, although no details of this of the rating procedures were given. Inter-rater reliability was consistently high for all investigations and varied little according to the students' curricular experience. Inter-task reliability was difficult to attain, since some students performed well on one task but poorly on another. For all investigations, mean performance was higher for students from the 'hands-on' science district than from the 'textbook' science district. The correlations between investigations and standardised multiple-choice tests were only moderate in magnitude, suggesting that these tests measured different aspects of science achievement.

For the other forms of assessment, notebooks provided the closest approximation in reliability and validity. The next closest surrogate for observed investigations were the computer simulations. Mean performance was comparable to the 'benchmarks', as were the patterns of correlations. However some students who scored high on the benchmarks scored low on the computer simulations and vice versa. The paper and pencil surrogates did not fare as well. Compared with the benchmark observed investigations, the short-answer items were less reliable and correlations with the standardised achievement test and aptitude test were higher. Moreover, mean performance of students experienced in hands-on science did not differ significantly from the performance of the students receiving 'textbook' science.

Among the authors' conclusions were that raters can reliably assess students' hands-on performance on complex tasks in real time. They considered that reliabilities are high enough that a single rater can provide a reliable score. The assessment of this hands-on performance can distinguish students with different instructional histories. However there is considerable variability according to the task and assessments that are closely linked to a specific domain of knowledge (e.g. electric circuits) are more sensitive than more general process assessment (e.g. paper towels). Notebooks and computer simulations can serve as surrogates for actual investigations.

Summary of main points from studies of reliability of assessment based on teachers' judgements

Evidence of high weight

- The reliability of portfolio assessment where tasks were not closely specified was low (Koretz *et al.,* 1994; Shapley and Bush, 1999). This finding has been used as an argument for increasing the match between task and assessment criteria by closer specification of tasks.
- The finer specification of criteria, describing progressive level of competency, has been shown to be capable of supporting reliable TA while allowing evidence to be used from the full range of classroom work (Rowe and Hill, 1996).

- Studies of the NCA for students aged 6 and 7 in England and Wales in the early 1990s found considerable error and evidence of bias in relation to different groups of students (Shorrocks *et al.*, 1993; Thomas *et al.,* 1998).
- Study of the NCA for 11 year-olds in England and Wales in the later 1990s shows that results of TA and standard tasks agree to an extent consistent with the recognition that they assess similar but not identical achievements (Reeves *et al.,* 2001).
- The clearer teachers are about the goals of students' work, the more consistently they apply assessment criteria (Hargreaves *et al.*, 1996).
- When rating students' oral proficiency in a foreign language, teachers are consistently more lenient than moderators, but are able to place students in the same rank order as experienced examiners (Good, 1988a; Levine *et al.*, 1987).

Evidence of medium weight

- In interpreting correlations of TA and standard task results for seven year-olds, the variability in the administration of standard tasks should be taken into account. (Abbott *et al.*, 1994).
- Teachers who have participated in developing criteria are able to use them reliably in rating students' work (Frederiksen and White, 2004; Hargreaves *et al.*, 1996).
- Teachers are able to score hands-on science investigations and projects with high reliability using detailed scoring criteria (Frederiksen and White, 2004; Shavelson *et al.* 1992).

4.2.2.Evidence from studies concerned with validity of assessment by teachers

The 18 studies designated as being mainly concerned with validity include five providing evidence of high weight, eleven of medium weight and two of low weight. The studies are divided into three groups for this discussion:

- those in which there is evidence about the process of assessment and about influences on the outcome that might call into question what is being assessed by teachers; the concern in these studies is with the extent to which this an adequate measure of the skills and knowledge intended to be assessed (construct validity);
- those in which there is a comparison between TA and performance on another measure of the same type of achievement; in some cases the evidence relates to construct validity and in others to how well the TA correlates with another measure (concurrent validity);
- those in which there is a comparison between the teachers' predictions of how students will perform on a test, drawn from their observations and interactions with students, and the actual performance on that test (predictive validity).

Studies of the process of assessment by teachers

In this section, there were seven studies all relating to the construct validity of assessment by teachers. Table 4.4 lists them in the order in which they are discussed.

Table 4.4: Studies providing evidence about the process of assessment by teachers

Study	Age of learners (years)	Achievement assessed	Study type	Overall evidence weight
Bennett *et al*. (1993) Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill	5 - 8	Reading English Maths	Exploration of relationships	High
Hall *et al*. (1997) A study of teacher assessment at Key Stage 1	6 - 8	English Maths Science	Evaluation: Naturally-occurring	High
Hall and Harding (2002) Level descriptions and teacher assessment in England: towards a community of assessment practice	6 - 8	English Maths	Evaluation: Naturally-occurring	Medium
Gipps *et al*. (1996) Models of teacher assessment among primary teachers in England	10 - 12	English Maths Science	Description	Medium
Radnor (1995) Evaluation of Key Stage 3 Assessment in 1995 and 1996. Evaluation of Key Stage 3 assessment arrangements for 1995	13 - 15	English Maths Science	Evaluation: Naturally-occurring	Medium
Koretz *et al*. (1994) The Vermont Portfolio Assessment Program: findings and implications	5 - 16	Reading Maths	Evaluation: Naturally-occurring	High
Shapley and Bush (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience	3 - 8	Reading Writing Maths	Evaluation: Naturally-occurring	High

The study by Bennett *et al.* (1993) involved a total of 794 students in 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 academic test was the Einstein Assessment of School-Related Skills, the behaviour grades were given by the teachers and the academic performance was derived from grades given by the teacher and from ratings of basic skills. The ratings were guided by a short verbal description so that 'in contrast to grades, the skill areas rated were common across school districts and were more specifically defined, making confusion with unrelated classroom behaviours less likely' (Bennett *et al.,* 1993, p 349). Regression analysis was used to test a model of influence of each variable on TA of academic achievement (ratings and grades) when other variables were held constant. Regression analyses were run separately for each grade level within each district, thus permitting both location and grade to be treated as replications.

The results were 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. Also, test scores were consistently and significantly affected by academic grades (less so for academic ratings). Grades 1 and 2 had similar patterns to each other but differed in some respects from Kg. Behaviour grade had a consistent direct effect on academic judgement (after controlling for gender, academic score and missing data). Academic test scores showed a similar relationship with both academic grade and academic rating. Only the indirect path beginning with gender (through behaviour grade to academic judgement) had consistent effects. The authors report that 'these indirect effects 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' (Bennett *et al.*, 1993, p 350).

Bennett *et al.* (1993, p 351) conclude as follows:

> '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.'

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 kindergarten, 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 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).

Hall *et al.* (1997) described their study of TA of seven year-olds in the National Curriculum Assessment in England and Wales as exploratory, designed to describe and explain rather than to predict or generalise. However, the study collected data from a sample of 45 schools, varying in size, socio-economic background, urban-rural mix and denomination. Data were collected by semi-structured interviews held after the teachers had completed their TA and standard tasks in 1993. Documentary evidence in the form of policy statements and samples of children's work and reports was also collected. The interviews were semi-structured, in order to give teachers

the chance to ‘reveal their own attitudes to and understandings of TA and the strategies they use to assess their students’ (Hall *et al.,* 1997, p 108).

Qualitative analysis of the data identified a 'model' of conducting TA which the authors claimed was the only one fitting the interview data. Quantitative aggregation of data (frequencies of response) was used to indicate the incidence of various practices. The model identified a series of stages in how teacher conduct TA. These were:

1. Assessment planning stage
2. Observation stage
   Strategies used are mainly observation, questioning and discussion. Only a minority of teachers use previous records. In the ongoing process of gathering information, these same methods predominate, but there is a wide range of strategies used, including conferencing, and the use of optional standard tasks, and teacher-devised tasks.
3. Specific task stage, where teachers are concerned to match work to individual needs. Differentiation by outcome begins to give way to differentiation by task (Hall *et al.*, 1997, p 110).
4. Continuous review stage
   This stage is recursive with stage 3 in that judgements made at stage 4 inform the allocation of work. A characteristic of stage 4 is ‘that it is now largely, though not wholly, a formalised process of assessing the extent to which attainment targets have been attained' (Hall *et al.*, 1997, 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 fourth stage is the longest, when the teacher not only gathers evidence fairly systematically but makes judgements about it. (It may signal a gradual change from a formative purpose to a summative one.)
5. Levelling stage
   This 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.

The study reported concern among teachers about making fair and accurate assessment. Particular attention was paid to the assessment of process skills; this provided the greatest challenge to teachers. It was in this area that teachers used 'intuition' rather than more systematic data and interpretation. The authors also report a sense of ‘professional mistrust’ amongst teachers, who treated with caution – and even suspicion – the assessments of other teachers (Hall *et al.*, 1997, p 113). A minority of teachers referred to assessing the broader aspect of children's learning beyond the national curriculum requirements.

Although the teachers were reported to be unanimous in their claim that the need to assess caused them to plan in greater depth, it also caused them to concentrate more on curriculum coverage rather than to follow their own or children's inclinations and interests. This suggests that the change might have been in detail of planning, rather than in depth. The teachers were uncertain about how and how often evidence

of progress should be documented. Over three-quarters reported using record sheets and checklists.

The authors conclude that the most significant aspect of the model is that TA is seen as an activity which influences all aspects of their work, 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 these teachers were integrating assessment into teaching and not merely adding 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, not all impact was positive – for example, focusing the assessment on a single year (Year 2) rather than the whole key stage.

A later study by Hall and Harding (2002) of year 2 teachers, providing medium weight evidence for the review, was conducted after the introduction of level descriptions in the National Curriculum Assessment of England and Wales. Level descriptions (LDs) for each achievement level replaced a series of separate 'statements of attainment' and were intended to be used holistically through a process of judging 'best fit' of evidence against a level with cross-checking against adjacent levels. The purpose of Hall and Harding's study was to explore the extent to which, in using these descriptions, teachers were collaborating in a 'community of practice' to develop a shared understanding of their meaning in practice. The authors argue that this was particularly necessary, given that the descriptions 'comprise mixtures of concrete and specific elements, as well as abstract and general elements, and they exhibit little intrinsic coherence, thus making it difficult for teachers to interpret and apply them consistently' (Hall and Harding, 2002, p 2).

Hall and Harding studied the procedures of teachers in six schools in 1998 and 1999. They tape-recorded and transcribed interviews with the teachers of seven year-olds and the assessment coordinators in all six schools and the LEA advisers of assessment. They also collected from the schools documentary evidence (such as portfolios, record sheets, and school and LEA assessment documents) and they observed one assessment meeting in one of the schools. A grounded process of clustering and categorising was used to identify patterns in the data, which were also informed by themes from the literature.

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, in the latter, teachers 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. In brief, collaborative schools showed compliance and acceptance of goals (contrasted with reluctant compliance and resistance for individualistic schools); 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 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 teacher showed that they had limited access to this expertise - for a variety of reasons, some relating to the large number of initiative which resulted in TA being put 'on the back burner'. As a result, teachers were left to depend mainly on one another for support. The researchers noted a decline between 1998 and 1999 in the level of collaboration within the schools and in neither year was there any collaboration between schools, although this had been a feature in all six schools in the early days of NCA. They also found that the potential for both learners themselves and their parents to be more actively involved has not been fully explored and exploited. The divergence in assessment practice among individual teachers working in isolation would be expected to lead to differing interpretations of criteria and hence to reduced reliability and validity of the assessment.

Hall and Harding (2002) concluded, *inter alia*, that the lack of funding for teachers to moderate their TA results served to tell teachers that the results of the external testing programme were prioritised over TA. Further '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, p 12). 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 national curriculum assessment.

Gipps *et al.* (1996) conducted a study of the assessment practice of teachers of 11 year-olds (Year 6) in the year 1994-95, just before the introduction of level descriptions in the National Curriculum Assessment of England and Wales. Their procedures followed similar lines to their earlier study of the TA practice of teachers of children aged 7 (Year 2) in 1992 (McCallum *et al.*, 1993). In a similar manner to Hall *et al.* (1997), they sought to identify styles of practice. For the 1996 study, teachers were interviewed using a technique of 'quote sort'. This involved reading a series of quotes about assessment practices, collecting evidence, making decisions about NC levels and recording, which had been developed through earlier interviews with Year 6 teachers. Each teacher was asked to decide whether a quote was saying something that was 'very like me', 'quite like me', 'not really like me' or 'not at all like me'. The teachers were also asked to talk about their reasons for their selections in a 'diagnostic debriefing session'. The teachers were observed for a morning. In addition, five teachers, who were very different in their approaches, were observed over four days. Qualitative analysis of the data was carried out, using a form of constant comparison of teacher responses to emerging clusters, using information from the observations and from case study data from the schools.

Four clusters of teachers emerged. Using a similar technique for the Y2 teachers, three groups had been identified (McCallum *et al.*, 1993). Only one of these, the 'systematic planners', can easily be related to the model reported by Hall *et al.*, (1997). The four emerging models for Y6 were described in terms of teachers who were dubbed:

- Testers: 11 of the 29 teachers
- Frequent checkers: five of the 29 teachers
- Markers: eight of the 29 teachers
- Diagnostic trackers: four of the 29 teachers

The characteristics of these are summarised as follows.

Testers do more than talking, listening and note-taking during normal activities; they plan assessment and give special tasks. They refer to levelled tasks when assigning levels; assessment is essentially 'bolt on'.

Frequent checkers also plan assessment tasks to be carried out at various times during the year, but also give more short informal tests of spelling and tables, more self-designed assessment tasks (aimed at groups, year groups, or sets). They also 'eavesdrop' and talk to children to pick up misunderstandings which are noted and used to inform the next day's or week's planning for teaching (but not on an individual basis). They do not like testing and data collection is an unobtrusive activity in most cases; children's performance on the small tasks or in daily activities becomes the evidence of attainment and recording a level is done more frequently than half-termly.

Markers have teaching as their priority. They use marking schemes which later need to be converted into NC levels; work is aimed at the whole class and regular work is used rather than assessment tasks/material as evidence for assigning levels; they see marking as assessment; the work is loosely based on the NC; they are not interested in taking notes of observations and rely a lot on memory; they do not record NC attainment as they go along but convert marks from their personal making scheme into levels for the school records half-termly or termly.

Diagnostic trackers are characterised by detailed planning for different NC levels, day-to-day tracking of children as they cope with the work, and TA that uses techniques of research - questioning, observation and recording incidents as they happen. They integrate assessment with teaching and they assign levels by the 'best fit' model based on the everyday work the children have done.

The authors are not able to make conclusions about the 'accuracy' of TA judgements made in these different ways. They say 'it may differ, or it may not' (Gipps *et al.*, 1996, p 181). However, some of their observations, made in the context of comparing the Y6 results with those from their previous work with Y2 teachers, suggest that there is variation between teachers which would be expected to make some TA more consistent with national curriculum criteria than others. For example, they found the following:

- Informal and 'qualitative' approaches to assessment, while more evident at the age of seven, are nevertheless a key feature at the age of 11.
- At both ages (seven and 11), some teacher do not adopt the use of NC levels but rely on personal criteria.
- At both ages, some teacher collect large quantities of evidence to support their assessment.
- At both ages some teachers are very systematic in their planning and assessment practice.

The study reported by Radnor (1995) was an extensive evaluation of the national curriculum assessment arrangements and procedures in England and Wales for Key Stage 3 (14 year-olds) in English, mathematics and science in 1995 and 1996. Data were collected through visits on two occasions to 39 schools, which constituted a 'core group'. The visits to the core group informed the development of questionnaires which were sent to 317 schools spread across the regions of England and Wales. The core group also provided students' marked scripts (about 2,000) which were scrutinised for evidence of students' misunderstanding or mismarking. Data are extracted from only one part of this study, that concerning teacher assessment. It provides evidence of medium weight for the review.

The study reported that, in order to complete their TA, English teachers found written work completed in class most effective for gathering evidence. Mathematics and science teachers tended to rely on school examinations and tests. All teachers believed cross-moderation among teachers to be important, but the constraints of finance and time made this difficult. The tendency was, instead, for individual teachers to use standardised test material or work with standardised exemplar materials, such as those provided by SCAA/ACAC. As far as the relationship between TA and test results was concerned, teachers were divided into 'levellers' and 'differentialists'. Levellers expected the two to show comparable levels and they finalised their TA after the test results were considered. Differentialists did not expect a match, considering that TA and tests assess different things. They completed their TA without taking tests results into consideration. English teachers were mostly differentialists, while levellers predominated in the mathematics and science teachers.

The author does not draw conclusions from these reported findings. However, the findings that the mathematics and science teachers were basing their TA on what is presumed to be a narrow range of work casts some doubt on its validity as a reflection of students' achievement across the full range of the curriculum, including aspects that cannot be assessed by tests. Teacher who used tests for their TA might well expect that it should give comparable results to the national curriculum tests, while those using a wider range of evidence might expect differences. However there was no evidence in the report as the whether there was a relationship of this kind in the data.

Finally in this section, it is relevant to refer again to the studies of Koretz *et al.* (1994) and of Shapley and Bush (1999), which provide evidence in relation to validity as well as to reliability. The evidence is tentative on account of the low reliability of the portfolio measures and is thus judged to be of medium weight.

In relation to validity, Koretz *et al.* (1994) used the 'uniform tests' in writing (one single prompt) and mathematics (multiple-choice test) to explore the extent to which the Vermont portfolios assessed similar or different achievements to these tests. The unreliability of the portfolios, of course, made a direct comparison problematic. However, the authors used an approach that allowed a tentative indication of 'what relationship might have obtained had scoring been reliable' (Koretz *et al.*, 1994, p 10). The result was that the evidence pertaining to construct validity was not persuasive. In some respects, expected relationships were found: for instance, the correlations between the writing portfolio and the writing uniform prompt were

consistent with other research. But, in other cases, the relationship showed little evidence of validity. For example, the correlations between the mathematics portfolio score and the writing test scores were about the same as with the mathematics test scores. The authors point to other research leading to the same conclusions as to the validity of portfolios. They also note that:

> 'To examine convergent and divergent evidence, one needs clear definitions of the domains the assessment are designed to tap and adequate measures of related constructs. In the case of the mathematics portfolio program, neither was available' (Koretz *et al.*, 1994, p 11).

As noted earlier (section 4.2.1), in their study of the Dallas reading and language arts portfolio assessment for young children, pre-Kg to Grade 2, Shapley and Bush, (1999) reported that many of portfolios were inadequately completed by teachers. Validity was compromised by teachers not adhering to work sample selection guidelines. To explore criterion-related validity, scores of the relevant students on the Iowa Test of Basic Skills (ITBS) were collected. Again, as for the Koretz *et al.* (1994) study, the low inter-rater reliability and limited degree of content coverage meant that investigation of construct validity could only be exploratory. For the Kg students, correlations between goal ratings and ITBS sub-tests scores showed low positive values. For grades 1 and 2 the correlations were somewhat higher, but in all cases there were no definite patterns indicating divergence or convergence. These findings do not necessarily indicate low validity, since 'it is possible that portfolio assessment, stressing both product and process, measures aspects of reading that standardised assessment do not measure. Divergent relations between ratings and mathematics scores were not firmly established. The differences between the mathematics sub-test score associations pointed to confounding effects of reading and language development when mathematics tests required students to read and solve written problems and to interpret data' (Shapley and Bush, 1999, p 126).

Studies of the relationship between teacher-assessed performance and performance on other measures of related but not identical performance

In this group, there were seven studies: two providing evidence of high weight for the review, three medium weight and two low weight. Table 4.5 gives information about the age group and achievement assessed, the study type and overall evidence weight in the order in which they are discussed. The pattern of discussing the high weight evidence first is interrupted on account of the similar types of achievement studied in Brown (1998) and Brown *et al.* (1996).

Table 4.5: Studies providing evidence of the relationship between achievement assessed by teachers and by other measures of related but not identical performance

Study	Age of learners (years)	Achievement assessed	Study type	Overall evidence weight
Brown *et al*. (1998) An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England	17 - 19	Science Practical maths/science/ tech	Exploration of relationships	High
Brown *et al*. (1996) The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations	17 - 19	Practical maths/science/ tech	Exploration of relationships	Medium
Meisels *et al*. (2001) Trusting teachers' judgements: a validity study of curriculum-embedded performance assessment in kindergarten to Grade 3	3 - 9	Reading Writing Maths	Exploration of relationships	High
Hopkins *et al*. (1985) The concurrent validity of standardised achievement tests by content area using teachers' ratings as criteria	9 - 11	Reading Writing Maths Science Social studies	Exploration of relationships	Medium
Sharpley and Edgar E (1986) Teachers' ratings vs standardised tests: an empirical investigation of agreement between two indices of achievement	6 - 10	Reading Maths General attitude Verbal intelligence	Exploration of relationships	Medium
Chen and Ehrenberg (1993) Test scores, homework, aspirations and teachers' grades	11 - 13	Reading Maths Bible	Exploration of relationships	Low
Papas and Psacharopoulos (1993) Student selection for higher education: the relationship between internal and external marks	17 - 20	Science Law Economics	Exploration of relationships	Low

The study of A-level performance in biology by Brown (1998) provides high-weight evidence of the effect of changing from external assessment to teacher assessment in the project component of the examination. Two data sets were drawn for the analysis: one from the 1993 examination, when the project was externally assessed by the Examination Board, and the other from that of 1996, when the project was assessed by teachers. Both samples were roughly representative of the subject entry in type of school, size of entry and geographical location. They constituted about 10% of the total entry. The input data consisted of candidates' scores on each of the theory components, on 13 process sections of the project for the 1993 data and, for

the 1996 data, on four teacher-assessed (TA) scores on the project (planning, implementing, interpreting and concluding, and researching). In both years, teachers were also required to make an assessment of candidates' practical skills during the course of laboratory work.

Construct validity was determined by correlational and factor analysis (principal component with rotation to varimax criterion) of candidates' scores. For the 1993 data, mean inter-correlation between the three theory papers for 1993 was 0.78, while inter-correlations between the theory papers and the project was 0.45, suggesting that something quite different was assessed by the project in comparison with theory papers. Inter-correlation between the project section scores and all other components were all positive and varied widely. Factor analysis showed a clear theory factor with very low loading of the project components. Teacher-assessed practical skills formed a separate factor.

For the 1996 examination, the mean inter-correlation among the theory papers was 0.84 and between theory papers and project 0.55. Factor analysis found two factors. Factor 1 was a theory factor but also two of the teacher assessed project skills loaded significantly onto it. Also, the teacher-assessed practical skills loadings were different from the 1993 findings. The author suggests that this was evidence that the TA procedure no longer assessed a construct different from theory and the project.

Brown (1998) concluded that the 1993 data showed that, 'overall, the project demonstrated construct validity in that it tested something that was different from the objectives tested by the theory papers… Of considerable interest was that the project factors received very low loading from the scores on practical skills derived from continuous assessment over the duration of the course' (Brown, 1998, p 94). He concluded that the two forms of practical - the ongoing skills during the course and the project - tested different constructs from each other and from the theory papers. For the 1996 examination, however, Brown suggested that the evidence for construct validity is much less compelling. He noted that, in 1996, teachers were required to assess four skills in the project, two of these being the same skills assessed on a minimum of two occasions as part of the ongoing practical/lab work. The outcome of these requirements was that the assessments in 1996 became more similar to the theory assessment than they had been in 1993. 'A speculative suggestion to explain this might be that a halo effect operated. Having assessed the practical abilities of their candidates over two years ...why would they expect different performances to emerge from the same candidates on the project?' (Brown, 1998, p 94).

In an earlier study, providing evidence of medium weight, Brown *et al.* (1996) investigated the construct validity of teacher assessment of practical skills in science. They used data from A-level candidates theory papers and from the teacher assessed scores on practical skills. Samples were selected of candidates for biology, physics and chemistry 'A' level examinations; the exact year of the examination is not reported. For investigating construct validity, inter-correlations among theory and practical components were computed separately for each science group and the same data also subject to factor analysis. The teacher assessment of practical skills was carried out by teachers for their own students during their 'A' level courses on two occasions. The skills specified were three for biology (A, B and C), four for chemistry (A – D) and five for physics (A – E).

The authors found that inter-correlations between practical skills were lower than between theory exams, varying with the science subject. In biology and physics, mean inter-correlations between theory and practical skills were lower than the inter-correlations of the individual tests. In chemistry, the inter-correlation between the theory and practical skills was higher than the inter-correlation of the practical skill results alone. The authors say that this suggests lower construct validity of the practical skill assessment in chemistry.

Factor analysis showed that two practical skills (A and B) consistently load higher onto a different factor from the theory assessment scores. Practical skills C and D tended to load equally onto the same factor scores as the theory tests and the practical tests. The authors argue that this suggests that there is some evidence of a practical construct being tested by each of the skills A (using and organising procedures and materials) and B (observing, measuring and recording), but less so for the other skills. Factor analysis of the skills scores showed the skills tests in each of the subjects load onto two different factors. The authors suggest that the weight of evidence is that both groups of practical skills are strongly related to the science subject.

The authors conclude that teacher assessment of students' practical work seems to make a valid contribution to assessment in these subjects. There is less evidence, however, that different facets of practical work make extensive contributions. There may be several reasons for this. ‘Teacher assessment may be subject to a halo effect in which an over-arching impression of a candidate's quality leads to similar judgements being made of performance in the different skills. Or it is possible that the skills which have been identified for each subject are intrinsically interrelated and scores on them will inevitably be highly correlated' (Brown *et al.*, 1996, p 388). There was little evidence of the generalisability of skills assessment across subjects, indicating that the skills assessment was context dependent. They conclude that this suggests that what was being assessed were subject rather than skill domains and that the results indicate a need for continued assessment of skills within subject domains.

The validity of using a detailed system of checklists to assess much younger children was investigated by Meisels *et al.* (2001). The tool used by teachers was called the work sample system (WSS). WSS requires three forms of documentation; checklists, portfolios and summary reports. The WSS checklist comprises skills and behaviours presented in the form of a one-sentence performance indicator. The items in the checklists (differing for each grade) measure seven domains of development: personal and social; language and literacy; mathematical thinking; scientific thinking; social studies; the arts; and physical development. Teachers rate students' performance on each item of the checklist three times per year and compare the rating with national standards for children of the same grade. Teachers use a modified mastery scale: 1= not yet; 2=in process; 3=proficient. During these periods, the teacher also completes the summary report, summarising each child's performance in the seven domains and rating it as 1= ‘as expected’, or 2= ‘other than expected’ and compare it with past performance. Portfolios are used to collect work that illustrates students’ achievements, efforts and progress. Meisels *et al.* (2001) compared results from the use of the WSS with the result of the Woodcock-Johnson Psycho-educational Battery - Revised (WJ-R), administered by trained testers on two occasions during the year.

Correlations were computed between students' standard scores on the sub-tests of the WJ-R and the WSS checklist and summary report ratings of student achievement within the corresponding WSS domains. Three-quarters of the correlations between WJ-R and WSS were within the range 0.50 to 0.75 and 48 of the 52 correlations between the WSS and the comprehensive scores of children's achievements fell within the 'moderate to high range'. The authors' judged that correlations of 0.70 to 0.75 are optimal, since the two measures are of overlapping but not identical ranges of attainment and thus these findings were taken to indicate 'strong *prima facie* evidence for the concurrent aspects of WSS's validity' (Meisels *et al.*, 2001, p 84).

Four-step hierarchical regressions were carried out to examine the factors that accounted for the variance in students' spring WJ-R scores. The results indicated that significant associations between WSS spring ratings and WJ-R spring outcomes remained even after controlling for the potential effects of age, SES, ethnicity, and students' initial performance level on the WJ-R in literacy (Kg-2) and in mathematics (Kg-1). There were some differences across grades. In the first step of the regressions, only the demographic variables were entered. This model was significant only in Kg and second grade for language and literacy, and in Kg for mathematics. When entered into the second step of the regressions with the demographic variables, the WSS checklist was significant at all grades levels for both mathematics and literacy. It explained more than half of the variance in literacy scores in grades 1 and 3. When the summary report was entered in the third step, both the summary report and the checklists contributed significantly in explaining the variance in the spring WJ-R literacy score for Kg-2. In the third grade, the checklist alone was a significant predictor of the language and literacy score. In mathematics, the WSS variables were significant predictors in step 3 of the regressions for Kg-3. Analysis of the mean WSS scores for the third grade indicated that teachers overestimated student ability on the summary report compared with the WJ-R.

The authors concluded that the results of the correlational analyses provided evidence for aspects of the validity of the WSS. 'The WSS demonstrates overlap with the standardised criterion measure and makes a unique contribution to the measurement of students' achievement beyond that captured by WJ-R test scores. The majority of the correlations between the WSS and the comprehensive scores of children's achievements (broad reading, broad writing, language and literacy and broad mathematics) are similar to correlations between the WJ-R and other standardised tests' (Meisels *et al.*, 2001, p 89). (The authors quote correlations between the WJ-R and other reading measures of 0.63 to 0.86).

'Overall the regression results provide evidence that WSS ratings demonstrate strong evidence for concurrent aspects of validity, especially regarding students' literacy achievement.' (Meisels *et al.*, 2001, p 90). They also note that the WSS discriminates accurately between children who are/are not at risk. The authors claim that the results demonstrate that 'We can trust teachers' judgements of student performance when they rely on WSS' (Meisels *et al.*, 2001, p 91).

A difficulty in establishing validity by correlation between two scores when the reliability of one or other (or both) may be uncertain has been noted earlier, from the work of Koretz *et al.* (1994) and Shapley and Bush (1999). Commonly, when the result of assessment by teachers and of standardised tests are correlated, the

standardised test is taken as the benchmark or criterion. However, in the study by Hopkins *et al.* (1985), the reverse was the case. These researchers used teachers' ratings and rankings of students' achievement as the criteria for evaluating the concurrent validity of a battery of standardised achievement tests. Their aim was to test the claim that multiple-choice tests are not valid tests of important curricular objectives. This study provided medium-weight evidence for the review.

Teachers were individually interviewed by an evaluation specialist from the district office of testing. Standard interview instructions were followed in which teachers were asked to rate on a five points scale, their students' achievement in reading, mathematics, social studies, language arts, and science. Some of the teachers were also asked to rank the students in reading from best to poorest. Tests in each of these curricular areas (the Comprehensive Tests of Basis Skills, CTBS, form S, Level 2) were administered about two weeks later as part of the school district's annual standardised testing programme. Pearson correlations between students' raw scores on each of the five CTBS tests and the corresponding teachers' ratings (and ranking in the case of reading) were computed. Correlation coefficients were transformed to Z-coefficients. Analysis of variance was used to explore differences across teachers and content areas.

The mean Fisher Z-coefficients for each of the content areas was 'quite high'; mean average within-class correlation coefficients ranged from 0.74 for language arts to 0.60 for science. There was a considerable variation across teachers, with the average across all five areas ranging from 0.44 to 0.88. There were also significant differences among the five content areas, the validity coefficients for language arts, reading and mathematics being significantly higher than for science and social studies. However, the authors note that some of these differences could have been due to test-length since the science and social studies tests have far fewer items and consequently lower reliability than the other three tests.

The relationship between the ratings and rankings for reading were used as a test of reliability of the teacher judgements. The correlation was 0.85. Rankings were found to correlate significantly more highly with the standardised reading tests than the ratings. 'The superiority of the normalized ranks appeared to result primarily from the reluctance of some teachers to use the full range of ratings, masking true, but discernible individual differences' (Hopkins *et al.,* 1985, p 181).

The authors concluded that, 'The high degree of correspondence between standardised achievement test score and teacher judgements, especially in language arts, reading and mathematics, demonstrates that both have substantial validity (or less likely, that both have little validity). Because most standardised achievement tests from the major test publishers inter-correlate highly, these results for the CBTS probably differ little from the correlations that would have been realized had a different standardised achievement test battery been used' (Hopkins *et al.,* 1985, p 181). Consequently they claim that 'in general, standardised achievement tests have substantial validity'. Although the within-class validity coefficients were slightly higher for rankings than for ratings, the differences were small. Ratings can be obtained much more quickly and with less rater frustration; this appears to be a satisfactory way of obtaining TAs.

The findings of Hopkins *et al.* (1985) conflict with those of Sharpley and Edgar (1986), a study also providing evidence of medium weight. Carried out at about the same time with students of similar age to those in the Hopkins *et al.* study, Sharpley and Edgar also correlated teachers' ratings and standardised tests. However, their study was 'designed to evaluate the accuracy of teachers' ratings of reading vocabulary, reading comprehension, mathematics, and verbal intelligence when compared with standardised test scores; to explore if teachers' attitudes towards students biased their evaluations of those students' academic progress; and to determine if bias against boys existed in teachers' evaluations' (Sharpley and Edgar, 1986, p 107).

Using a standard rating form with a table of ratings from 1 to 5 (1 = high, 5 = low), teachers rated each child in their classes according to their general attitude and present level of achievement in the four areas. Progressive Achievement Tests (PAT) were used to assess each child in reading vocabulary, reading comprehension, and mathematics. The Peabody Picture Vocabulary Test- Revised (PPVT-R) was used to assess receptive vocabulary. Mean ratings, standard test score and correlations were computed separately for boys and girls.

Correlations between ratings and test scores were significant at the 0.01 level in 19 out of the 20 correlations for boys and 15 out of the 20 correlations for girls. However, although significant, they were low and even the highest (0.56 for girls comprehension scores and ratings) accounted for only 31% of the variance. Results for verbal intelligence were particularly low, at 0.41 for boys and 0.15 for girls, leading the author to conclude that there is no evidence that teachers can accurately assess a child's verbal intelligence.

The teachers' ratings for girls, on the five-point scale, showed higher ratings for general attitude and verbal intelligence than for reading and mathematics. For boys, verbal intelligence was rated highest and general attitude lowest of the five ratings. Ratings on comprehension and vocabulary were significantly associated as were comprehension and mathematics. General attitude was not significantly correlated with ratings on the other variables. The standardised tests and the teachers' ratings showed girls scoring higher than boys on vocabulary, reading comprehension and mathematics, but not on verbal intelligence, where the difference was negligible. Teachers were no more accurate in assessing boys or girls

Given that the aim was to examine the accuracy of teachers' ratings when compared with standardised test results, the correlations do not support a strong commonality between the two assessment procedures. Thus, in contrast with Hopkins *et al.*, (1985) these authors conclude that there is little educational significance in the low correlations: 'It appears from the data, that although there are significant correlations between teacher ratings and standardised test scores, there is little direct meaning in these results' (Sharpely and Edgar, 1986, p 110). The authors suggest that the teachers' ratings and test scores assess two 'non-equivalent domains'.

The study by Chen and Ehrenberg (1993), investigating whether TA are vulnerable to influence from non-learning factors (such as socio-economic status, homework and gender) provides evidence of low weight, mainly on account of doubts of the reviewers concerning the analysis, but also the low relevance of the conditions of the study to the aims of this review. The study was conducted on a heterogeneous

sample of sixth-grade students in a medium-sized, affluent town in Israel. Information was collected about the students' background, achievement on standardised multiple-choice tests based on the formal school curriculum, students' aspirations, regularity in preparation of homework, as rated by the teachers, and teachers' grades. An over-identified path model relating these variables was tested through path analysis.

The study found that the strongest direct influence on teachers' grades was homework. Achievement scores was the next strong influence on teachers' grades. There was a very small, but statistically significant, influence of aspirations: those students with higher aspirations receive slightly higher evaluations as a result of these aspirations alone (all other variables being controlled). The authors concluded that the study confirms the hypothesis regarding the over-identified path model, which explains 72% of the variance in the teachers' grades. 'It suggests that teachers grade their student properly according to their knowledge of the subject matter, their efforts and aspirations only. There is no indication that they take into consideration irrelevant factors like gender or SES. However the excessive importance attributed to homework indicates indirect preference for students of high SES and female students (since homework is influenced by SES, aspirations and gender). The correlation between grades and homework may be somewhat inflated but it still suggests a much higher relationship than expected' (Chen and Ehrenberg, 1993, p 414).

The final study in this group was also one in which two measures were compared in the context of entrance to higher education. It provides evidence of low weight for the review. The study by Papas and Psacharaopoulos (1993) investigated, in the context of education in Greece, the correlation between internal marks and the marks gained on an external examination for university entrance. Until 1988, a combination of these two kinds of marks was used for university entrance decisions. The hypothesis tested was that, if these are highly correlated, then there may no need for external examinations.

The authors found that, in terms of mean internal and external marks, there is a substantial difference by school type and subject cluster. The match between external and internal marks was closer for those in private and selective schools, and for those aspiring to enter medical or law school. However, 'all correlations are on the high side revealing that the external marks somehow validate internal school marks.' (Papas and Psacharaopoulos, 1993, p 400). The authors conclude that on account of the high concurrent validity of the internal marks, external examinations may not be necessary since 'the marks in the last three grades of secondary school seem to reflect fairly well how a student will perform in the external examination' (Papas and Psacharaopoulos, 1993, 401).

Studies where there is a comparison between teachers' prediction of students' test performance and their actual performance on a specific test

Six studies were concerned with the predictive validity of assessment by teachers. One of these provided evidence of high weight and the other five provided medium-weight evidence. In each case, TAs of students' likely success in response to

specific items or on a tests as a whole were compared with their actual performance at a later date. Table 4.6 summarises information about the age groups and achievement involved, the study type and overall weight of evidence.

Table 4.6: Studies where assessment by teacher is used to predict achievement on a specific test

Study	Age of learners (years)	Achievement assessed	Study types	Overall evidence weight
Coladarci T (1986) Accuracy of teacher judgements of student responses to standardised test items	8 - 11	Reading Maths	Exploration of relationships	High
Wilson and Wright (1993) The predictive validity of student self-evaluations, teachers; assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachian schools	11 - 17	Maths Other Verbal reasoning	Exploration of relationships	Medium
Crawford L *et al.* (2001) Using oral reading rate to predict student performance on statewide achievement tests	5 - 10	Reading Maths	Exploration of relationships	Medium
Good and Cresswell (1988) Can teachers enter candidates appropriately for examinations involving differentiated papers?	15 - 16	Science History French	Exploration of relationships	Medium
Delap (1994) An investigation into the accuracy of A-level predicted grades	17 - 18	English Maths Science History Geography Sociology	Exploration of relationships	Medium
Delap (1995) Teachers' estimates of candidates' performances in public examinations	17 - 18	English Maths Science Sociology History Geography	Exploration of relationships	Medium

Studies in this section involved teachers in judging the extent to which their students would be able to succeed in specific tests or examinations, in some cases on individual items, in others on the test as a whole, but in the knowledge of what individual items demanded. In effect, the test items acted as very specific criteria

against which teachers judged the achievement of their students based on their observations and interactions during teaching.

In the only study in this section providing evidence of high weight, Coladarci (1986) asked teachers to assess success or failure on specific items of standardised tests. Teachers of third- and fifth-grade students gave their judgements of their students' achievement in interviews in which the interviewer randomly selected six students (two from each ability group). For each student, the interviewer asked the teacher to indicate whether he or she thought the students correctly answered specific items on the SRA test (Science Research Associates Achievement Series 1978). This was done for each item on the Reading Vocabulary, Reading Comprehension, Mathematics Concepts and Mathematics Computation sub-tests for the third and fifth grades (Form 1). The students had taken the test two weeks earlier, but the results were unknown to the teachers. Item-level results for both the tests and the teacher's judgements were summed to form sub-test results, total reading, total mathematics and total test results. Correlations were computed between each performance and judgement measure, and three measures of achievement: the actual achievement on the test items, the students' total score across the four sub-tests (i.e. total test) and the teachers' judgement of the student's general performance level ('below', 'at', or 'above' grade level).

Overall, the aggregate measures of teachers' judgements of their students' responses to items on a standardised achievement test correlated positively and substantially with aggregate measures of students' actual responses (coefficients ranging from 0.67 to 0.85). However, for all sub-tests, some students were judged correctly for fewer than half of the test items, whereas other students were judged correctly for nearly all the items. Analysis of variance indicated that the reason for this was a combination of teacher effect and student effect. There were significant individual differences among teachers in the accuracy of their judgements and overall teachers were least accurate in judging low-performing students and most accurate in judging high-performing students.

The author points out that the finding in relation to greater accuracy of teacher judgements for high-achieving students is expected. Teachers would be more likely to report that an 'above' grade level student would get the item correct and these students would be more likely to succeed. No simple response set would work for students further down the achievement scale. 'For the moderate and low-achieving student, teachers doubtless realised that there were many items that the student would not answer correctly. What was difficult was to decide where the errors would occur. And the lower the student's proficiency, the more difficult - and inaccurate - this judgement was. These results point tentatively to the disturbing implication that students who perhaps are in the greatest need of accurate appraisal made by the teacher in the interactive context are precisely those students whose cognition has a greater chance of being misjudged' (Coladarci, 1986, p 145).

The author suggests that the relationship between teacher accuracy and task can be explained, in part, by (a) the degree to which teachers provide direct instruction in the task domain and (b) the amount of information teachers have that bears on student proficiency in that domain. In mathematics, typically, there is more direct instruction in computation than in concepts. Another factor might be the complexity of the task: open-ended test items would be likely to aid understanding here rather than

the multiple-choice items of the SRA tests. Thus, Coladarci concludes that 'applied to the interactive decision making, these results suggest that the accuracy of a teacher's judgement is influenced by characteristics of the teacher, student and academic task' (Coladarci, 1986, p 146).

A study by Wilson and Wright (1993) was another investigation of correspondence between TAs and other measures of performance, but differed from those already discussed in the inclusion of students' self-assessments. The study provides evidence of medium weight for the review. The investigation involved 306 students, all attending one of the four rural secondary schools in Appalachia. Students in grades 8 – 12 were involved, but only in Grade 11 were there sufficient numbers for some analyses to carried out. Two measures of teacher assessments of students achievement were used. In one, teachers predicted students' performance on the Verbal Reasoning and Numerical Ability scales of the Differential Aptitude Test (DAT), before this was administered. In the other, teachers gave their rating of academic ability on a five-point Likert-type response scale. This rating was repeated four weeks later, after the DAT testing. After the administration of the DAT to the students, they were asked to respond to 'How well did you do on the DAT?'. This was repeated one to two weeks later as a retest. Student grades in English and mathematics were obtained from the school records. Test-retest estimates were calculated for the students' self-evaluations and the accuracy of self-evaluations was estimated by comparing actual performance with self-evaluations of the DAT scores. The relationship between students' self-evaluations, teacher assessment and course performance variables and actual performance on the DAT were investigated by correlation and multiple regression analyses (used only for the eleventh grade).

Overall, the strongest correlates with the criterion variables (verbal reasoning, VR, and numerical ability, NA) were the teachers' evaluations (either of probability of success or of academic ability) and students' self-estimates. Teacher assessments' estimates correlated with actual performance fairly stably across grades. The teachers' estimates were moderate predictors for both verbal and numerical ability (coefficients 0.59 for VR and 0.47 for NA). Eleventh-grade, multiple-regression results showed that three variables were effective predictors of VR scores: teacher assessments of academic ability, student self-evaluation and grade averages. All were significant predictors, with the self-assessment just slightly stronger. For numerical ability, two of these variables were the same but TA of the probability of success replaced TA of academic ability.

The authors cautiously concluded that 'based on the regression analyses for the 11th Grade sample, student self-assessments and teacher perceptions of student ability and probability of success are important moderately valid predictors of academic performance at least in an academic setting in which students are at risk of dropping out of school' (Wilson and Wright, 1993, p 268). For all grades, teacher assessments achieved moderate to strong correlations with student performance on the verbal ability test and slightly less strength for the numerical test. The authors suggest that these findings may indicate that teachers may rely upon a perceived 'verbal competency' dimension in judging a student's academic ability as well as for estimating a student's potential for success in completing a given course of study.

Crawford *et al.* (2001) investigated the predictive validity of a curriculum-based measurement (CBM) of reading aloud as an indicator of student progress in reading

and mathematics. The particular question addressed was 'How strong is the relationship between oral reading rate and future performance on state-wide reading and mathematic achievement tests?' Students were followed for two years, through grades 2 and 3, when they stayed with the same teacher. To assess reading rates, three passages were chosen for use during each year of the study. Each was modified to have approximately 200-250 words and to have cogent beginnings and ends. Passages were taken from the Houghton Mifflin Basal Reading series which was used in the school district and participating schools. Each passage was read aloud and timed for one minute; this was then scored for correct words and errors. In their third year, the students were tested on state-wide mathematics and reading assessments - both were criterion-referenced tests, containing multiple-choice questions and performance tasks.

The mean scores on the state-wide reading assessment were at the state-established target level in reading, but the mean scores on the state-wide mathematics assessment fell short of the established criterion. Sixty-five percent passed the reading assessment and 45% passed the mathematics assessment. The results for the oral reading showed a large increase in the number of correct words read per minute between second and third grade, but no relationship between initial reading rate and amount of gain.

Correlations between the timed oral reading rates and state scores (reading test) were moderate (0.60) for the third grade and slightly higher for the second grade (0.66). Correlations between oral reading rates and mathematics state score were a little lower and again slightly higher for the second (0.53) than for the third grade (0.46). When looking at the results for individual students who did and did not pass the state tests, it was found that, of 37 students with reading rates in the top three quartiles, 29 passed the state reading test, whereas only 29% of the students reading in the first quartile passed. Of the students reading at least 72 correct words per minute in the second grade, 100% passed the state-wide reading test in third Grade. A chi-square, representing second-grade reading rates and state-wide reading test scores, demonstrated statistical significance. Similar calculations for reading rates and the mathematics tests showed no statistical significance.

The authors conclude that 'longitudinal data presented in this study demonstrate that CBMs are sensitive enough to detect growth for almost every student, with 50 out of 51 students in this study improving their rate of reading over the course of one year. CBM procedures also seemed to lack bias, in that the gains students made on the measures were not an artefact of their starting points, as we found no significant differences between the amount of gain made by students who had low initial rates and those that had high initial rates' (Crawford *et al.*, 2001, p 320). The authors claim that the results demonstrate that teachers can rely on the accuracy of CBMs in monitoring the reading progress of all students regardless of skill level.

A study by Good and Cresswell (1988), providing evidence of medium weight for the review, was prompted by the introduction of the GCSE examination for 16 year-olds in England and Wales. This examination replaced the separate CSE and O-level examinations and, in order to accommodate a range of achievement levels, provided differentiated papers in certain subjects. Teachers were required to enter students for the appropriate paper and so had to predict their likely performance. The aims of the study were to explore the issues concerning the accuracy with which teachers

can predict examination performance (in order to enter candidates at the appropriate levels when differentiated papers exist) and to see whether there was any effect of the time at which the predictions were made. The study took place just before the introduction of the new examination and the students involved in the study took an experimental examination before taking the operational CSE or O-level.

The study collected and compared actual and teacher-predicted grades of candidates entered for the experimental GCSE examination in three subjects: history, physics and French. 'Descriptions of the examinations which included the ranges of grades available and specimen questions were provided to help teachers arrive at their entry decisions' (Good and Cresswell, 1988, p 291). Data relating to an operational CSE examination were also collected for a sub-sample of students for whom predictions were made at two different times, to evaluate the effect of the timing of the prediction. Frequencies were reported of predictions corresponding with actual grades exactly, or under- or over-predicted by one, two or three or more grades.

Good and Cresswell (1998) found, as in other studies where teachers predicted scores, that teachers were slightly more likely to over-predict than to under-predict. Forty percent of teachers' predictions were correct and a further 45% were only out by one grade. Between 2% and 3% of predictions were out by more than two grades. For the CSE predicted grades, where predictions were made in January and in May before the examination, the proportions of predicted grades were almost identical on the two occasions. Thus teachers were as able to predict grades early in the year as they were later in the year. The authors concluded from the results of this study that, in an experimental context, teachers are able to predict the probable achievements of their students sufficiently accurately to enter them appropriately (from a grade standpoint) for GCSE examinations using differentiated papers.

Two studies by Delap (1994 and 1995), investigated the accuracy of teachers' predictions of the examination grades of their students at 'A' level. These predictions are used in the UK as information for higher education institutions to use in selecting students to be given conditional offers of admission. In other countries, also, predicted grades are used in situation of high stakes for the students. Thus the predictive validity of these judgements is of considerable importance. Delap's work challenged the method used in previous studies to analyse data in determining the accuracy of predicted grades. The descriptive analyses previously used considered the effect of various variables (such as student age, ethnicity and gender, subject, examining board, etc.) on the difference between predicted and actual grades as if each variable was independent of others, ignoring possible interaction effects.

In Delap (1994), analyses were carried out for over 9,000 predicted 'A' level grades for about 3,000 students, collected in 1991 from the UCCA application forms on which teachers entered their predictions. There predictions and the grades actually obtained by the students were analysed first by the previously used methods and then by a two-level modelling procedure through which interaction effects between variables could be explored. The results of the first analyses were in many respects similar to the findings of previous studies. Mean difference between predicted and actual grades was approximately half a grade and there was a significant correlation between predicted and actual grades. The results of the two-level model were that almost all the variance between predicted and actual grades was attributable to the

applicant level variables with very weak effects at the subject level. There was a small but significant gender effect; predictions for males were more optimistic (less accurate) than those for females. Some evidence was also found to support the view that the predicted grades from further education establishments were more optimistic on average than those from other types of centre. Finally, the origin of the applicant had no influence upon the optimism or pessimism of teachers predicted grades. Delap (1994) concluded, 'It has been demonstrated that, unless the distribution of the actual grades for each of the sub-categories being compared are similar, the interpretation of the data will be misleading. For example, one may have been led to conclude from the summary analysis that there are significant differences between the accuracy of prediction for some of the examining boards. The finding from the multilevel analysis reveals that, while the difference exists in the raw data, once other factors are taken into account - ie gender, final grade, centre type and subject - there is no significant difference between the predictions for examining boards' (Delap, 1994, p 147).

In a later study, Delap (1995) analysed 'A' level predictions collected in 1992 for candidates from one examination board. Teachers were asked to use the scale of seven grades (A, B, C, D, E, N and U), as used by the examination board, in their estimations. The actual grades obtained were later collected. Multilevel analysis was used to analyse the data. The data were structured with candidate data at level 1 and school data at level 2. The function for the estimate of grade included actual grade, gender, etc.

From the distributions of estimated and actual performances, it was evident that the distributions were markedly different. 'For example, it is readily apparent that teachers were not inclined to provide estimates of low grades (N and U). Similarly more candidates obtained grade A than were estimated to do so' (Delap, 1995, p 79). The accuracy of teachers' judgements, indicated by the proportion of estimates that were accurate (the predicted grades were those obtained by the candidates), varied enormously across subjects and grade levels. For example, for physics, 84% of those estimated to gain A did so, while the proportion was only 18% for grade C. For chemistry, these figures were 28% for A and 27% for C. Overall the proportion of accurate grades was highest for mathematics and biology, and lowest for physics. The analysis of factors which influenced estimates showed that, for most subjects, there was a significant school effect, but no general trend relating to type of school; in three subjects (biology, geography and mathematics), there was evidence that teachers' estimates were slightly higher for females than for males; and the age of the candidates had a very small influence upon the estimated grade for only three of the eleven subjects. It was also found that when the estimates were made in the three to four months preceding the examinations, did not substantially affect the estimated grade.

Overall, Delap (1995, p 91) concluded that 'the teachers' estimates were not very accurate; about half of the estimates were optimistic and estimated grades of C, D or E were accurate on about one in four occasions'. He considered that, rather than to replace them, estimated grades can be of some value in providing information to complement decision-making processes used in higher education admission procedures.

Summary of main points from studies of validity of assessments based on teachers' judgement

Evidence of high weight

- Teachers' judgement of the academic performance of young children are influenced by the TA of their behaviour. This adversely affects the assessment of boys compared with girls (Bennett *et al.*, 1993).
- The introduction of TA as part of the national curriculum assessment initially had a beneficial effect on teachers' planning and was integrated into teaching (Hall *et al.*, 1997). Subsequently, however, in the later 1990s, there was a decline in earlier collaboration among teachers and sharing interpretations of criteria, as support for TA declined and the focus changed to other initiatives (Hall and Harding, 2002).
- The validity of a science project as part of 'A' level examinations for assessing skills different from those used in regular laboratory work was reduced when the project assessment was changed from external to internal by teachers (Brown, 1998).
- Teachers judgements guided by checklists and other materials in the work sampling system (WSS) were found to have high concurrent validity for assessment of Kg to Grade 3 students (Meisels *et al.*, 2001).
- Teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching (Coladarci, 1986).

Evidence of medium weight

- There is variation of practice among teachers in their approaches to TA, type of information used and application of national criteria (Gipps *et al.*, 1996; Radnor, 1995).
- There is conflicting evidence as to the relationship between teachers' ratings of students' achievement and standardised test score of the same achievement when the ratings are not based on specific criteria (Hopkins *et al.*, 1985; Sharpley and Edgar, 1986).
- The rate at which young children can read aloud is a valid curriculum-based measure of reading progress as measured by a standardised reading test (Crawford *et al.*, 2001).
- Tentative estimates of construct validity of portfolio assessment, derived from evidence of correlations of portfolios and tests, were low (Koretz *et al.*, 1994; Shapley and Bush, 1999).
- Teacher assessment of practical skills in science makes a valid contribution to assessment at 'A' level within each science subject but there is little evidence of generalisability of skills across subjects (Brown *et al.*, 1996).
- Teachers' perceptions of students' ability and probability of success on a test are moderately valid predictors of performance on the test, as are student self-assessments of their performance on a test after they have taken it (Wilson and Wright, 1993).

4.3 Synthesis of evidence: subsidiary review question

What conditions affect the reliability and validity of teachers' summative assessment?

Almost all the studies provide some evidence of conditions or variables that may influence the degree of reliability or validity of the assessment by teachers. Looking across the studies providing high- and medium-weight evidence, eight main factors emerge as the ones identified as having, or being likely to have, an impact on reliability and validity. The first four are sources of variation for which there is empirical evidence. The second four refer to those features of TA practice which have been identified by the study authors as likely to influence the accuracy of assessment by teachers.

- Variation linked to student variables (gender, age, special education needs, etc.)
- Variation linked to school variables, often unspecified
- Variation linked to the subject assessed
- Variation in procedures and evidence used
- The influence of the training of teachers in assessment procedures
- The influence of the specification of tasks in which students are assessed
- The influence of the specification of criteria and their meaning
- The influence of moderation and inter-teacher collaboration

The variation caused by these factors is important not only because they are sources of unreliability and thereby infringe validity, but because they may be sources of inequity or unfairness. Whether or not this is the case points to the complexity of the issue. For instance, when one sub-group appears to perform consistently lower on average than another in the same population, the reason may be a real difference in performance or bias in the assessment procedure (Gipps, 1994). Even when two different measures of the same achievement (such as TA and standard tasks or tests) show different patterns of sub-group performance, it is not easy to claim that one is more biased than the other. Moreover, the sources of differences between sub-groups are not independent of each other, but inter-related. There are many examples of this interaction in the following discussion. Thus, although the factors are discussed separately, their interaction has to be kept in mind.

4.3.1 Student variables

Several studies reported differences in TA related to the gender of students, which were not present in standard measures of performance. Reeves *et al.* (2001) and Bennett *et al.* (1993) found that, at the primary level, teachers under-rated boys. For upper secondary students, Delap (1995) also found a gender effect, with predicted grades for females being slightly higher than for males in mathematics, biology and geography. However, Delap (1994) reported little gender difference in the percentage of pessimistic, optimistic and accurate grades. Bennett *et al.* (1993) included teachers' perceptions of pupil behaviour in their model of influences on teachers' judgements of students' academic skill and found that this was a significant component in teachers' academic judgements. Since boys' behaviour was more

often perceived to be poorer than that of girls, boys' academic skills were perceived as being less adequate than girls'.

TA of pupils with special educational needs (SEN) was noted in three studies. Reeves *et al.* (2001), Thomas *et al.* (1998) and Shorrocks *et al.* (1993) all reported that TA results for students with SEN were frequently lower than their test results. Reeves *et al.* (2001) suggested that, rather than teachers under-estimating these students, the reason for the difference might be that the students were over-performing in the tests perhaps as a result of 'teaching to the test'.

The studies by Brown (1996 and 1998) of the assessment of science projects and laboratory skills indicated that teachers' judgements of these skills might be influenced by a 'halo' effect, the teachers' overall judgements of students' ability affecting their assessment of individual skills. Wilson and Wright (1993) suggested from their findings that teachers' judgements may be influenced by their perception of students' verbal competency. Similarly, Levine *et al.* (1987) reported an influence of overall grade on the assessment of oral language in the context of learning a foreign language.

4.3.2 Variation linked to schools

Much of the evidence in relation to this variable comes from studies of the implementation of National Curriculum Assessment in England and Wales. Reeves *et al.* (2001) found, across a random sample of schools, that a considerable amount of the variance in the difference between TA and test scores was due to the school. They found that this decreased over the years, indicating a gradual increase in uniformity of practice across schools. However, the reverse trend was hinted at by Hall and Harding (2002), who noted a decline in inter-school collaboration from the early to the later 1990s, with a divergence of practices in the implementation of TA in the early primary school. In particular, they noted a difference among schools in relation to involving students in self-assessment.

Thomas *et al.* (1998) reported an 'unexplained school level variation' in the relationship between TA and standard task results in the early 1990s. Some of this might well have been due to the differences in practice of TA observed by Gipps *et al.* (1996), by Hall *et al.* (1997), and by Hall and Harding (2002). However, explanations of the school variation in relation to the difference between TA and standard tests and tasks should take into account the findings of Abbott *et al.* (1994), who reported a considerable variation in several aspects of the administration of the standards tasks that would call into question the reliability of these measures, at least for those used in the early years of the national curriculum assessment, 1990-92.

4.3.3 Variation due to subject assessed

For assessment by teachers at the primary level, where the same teacher assesses English, mathematics and science in the national curriculum assessment, differences across these subject were reported by Reeves *et al.* (2001) and Shorrocks *et al.* (1993). Reeves *et al.* (2001) found that, for 11 year-old students, in mathematics, T'A' levels were higher than test levels, while in English and science the reverse was

the case. Consistency in direction and size of the difference were at a 'remarkably high level'. Shorrocks *et al.* (1993), looking at the assessment of six and seven year-olds, found close agreement between TA and standard tasks for English, but less in some aspects of mathematics and science.

At the secondary level, Radnor (1995) noted difference in the kind of evidence used in their TA by teachers of English, mathematics and science; English teachers made more use of classwork, and mathematics and science teachers relied more on school examinations and tests. Delap (1994) reported large differences in subjects in the proportion of 'A' level grades accurately predicted by teachers. The smallest variation was in French and the largest in history. Delap (1995) confirmed the subject variation and found no general trend across schools in the 'A' level estimates for different subjects. Levine *et al.* (1987) found that teachers over-rated performance of high school students in oral foreign language, but reported no difference in the patterns for French and Spanish.

4.3.4 Variation in procedures and evidence used

It is clear from the profiles of the studies in Appendix 4.1 that, in many cases, researchers depended on questionnaires or interviews with teachers which provided self-reports as to how their assessment were carried out. In other cases, no information was given about process. Thus, it is often uncertain as to what evidence was used by teachers, how it was gathered and how it was interpreted. In extreme cases, this may mean that the TA could be based on information picked up through daily interaction of teacher and students (more likely at the primary level where the same teacher covers all or most of the curriculum), or it may have been based on tests and formally marked work which is little different in its range of content and skill assessment from external tests and examinations (more likely at the secondary level, where teachers have less opportunity for extended interaction with each student).

In only a few studies were actual practices observed. As noted above, these studies reported considerable variation in practice among primary teachers in the early years of implementation of assessment by teachers as part of the national curriculum assessment in England and Wales (Gipps *et al.*, 1996; Hall *et al.*, 1997 and Hall and Harding, 2002). From interviews with secondary teachers, Radnor (1995) was able to probe practices and sources of evidence, finding different approaches used by teachers of English, mathematics and science. However, although it might be inferred that these differences would influence the reliability of the resulting measures, these studies provide no firm evidence on this point.

Koretz *et al.* (1994), and more particularly Shapley and Bush (1999), report considerable variation in teachers' implementation of portfolio assessment. The main shortcomings were in the failure to document the pieces of work in the portfolio and inadequate sampling of the range of work in the subjects assessed. However, the lack of reliability, as measured by inter-rated reliability, was thought to be due to insufficient specification of tasks to be included in the portfolios and inadequate training of the teachers. While there is no evidence of the effect of making these changes in the specific circumstances of the Vermont and Dallas portfolio systems, findings of other studies help to throw light on their likely influence on the reliability and validity of the assessments.

4.3.5 The influence of the training of teachers in assessment procedures

Different kinds of training are suggested as being likely to have a positive impact on the consistency of TAs. Koretz *et al.* (1994) and Shapley and Bush (1999) imply that teachers need more training in procedures developed for operationalising the portfolio system. Although Koretz *et al.* (1994) described the system as being developed 'bottom up', once the guidelines and scoring rubrics were identified, it was necessary for all teachers to adhere to them. The training was thus in practising how to conform to the agreed procedures. Levine *et al.* (1987) propose that training should be specifically focused on those aspects that research has shown to be sources of error. They note the influence of students' overall grade on the assessment of oral language and the consistent over-rating of student performance in foreign language oral examinations. They claim that these deficiencies would be susceptible to change by appropriately focused workshops.

Rowe and Hill (1996), Hargreaves *et al.* (1996) and Frederiksen and White (2004) all note that the participation of teachers in the development of criteria is an effective training in the reliable use of the emerging criteria. Although not all teachers can be involved in the development of criteria, this experience suggests that training should as far as possible aim to emulate such involvement and to give teachers a sense of ownership of the procedures and criteria to be used.

4.3.6 The influence of the specification of tasks in which students are assessed by teachers

There is evidence here from those studies where teachers made predictions of students' success on specific test items (e.g. Coladarci, 1986) or on tests as a whole (Wilson and Wright, 1993; Delap, 1994; Delap, 1995; Good and Cresswell, 1988). Colardarci (1986) found substantial agreement between predictions and actual performance, with variation across tasks that may be due to the amount of attention given by the teacher to particular types of task. Good and Cresswell (1988) concluded that teachers could predict with reasonable accuracy the success of their 16 year-old students in an examination, given information about what the examination assessed and examples of the items. The time at which they made their predictions, within one to four months, in relation to the examination made no difference. Delap (1995) was less optimistic about predictions of 'A' level success, finding systematic over-estimation and inaccurate predictions of middle-range grades.

Koretz *et al.* (1994) and Shapley and Bush (1999) conclude that greater standardisation of tasks to be placed in portfolios is needed to improve the reliability of the assessment, providing a better match between tasks and assessment rubrics.

4.3.7 The influence of the specification of criteria and their meaning

While the closer specification of *tasks* assessed, as advocated by Koretz *et al.* (1994) and by Shapley and Bush (1999), takes TA closer to a series of standard

tasks, Rowe and Hill (1996) provide evidence that the closer specification of *criteria* may be effective and preferable alternative. The subject profiles give detailed descriptions of what students can do at different points in development within each of various subject strands. The authors claim that these give meaning to each level used in reporting achievement and help teachers to understand the language used to describe students' progress. The authors report that the descriptors at each level were used consistently by different teachers, while allowing evidence to be used from across the range of work in each classroom. Meisels *et al.* (2001) also found that the use of checklists, describing skills and behaviours, enables teachers to make valid assessments of students language and mathematics achievement, using curriculum embedded evidence.

However, it is important to note that, as Frederiksen and White (2004) point out, the credibility of judgements of performance depends on the nature of the tasks and not just on inter-rater agreement. Thus, even though teachers may be able to use the subject profile indicators reliably, it is important that the assessment are of tasks that are valid in terms of the curriculum guidelines and provide worthwhile challenges to students. Further, as already noted, Coladarci (1986) reported that teachers are less reliable in assessing aspects which are not strongly represented in their teaching.

4.3.8 The influence of moderation and inter-teacher collaboration

Reference is made in several of the studies to various forms of moderation, including the adjustment of marks (Good, 1988), agreement across teachers (Radnor, 1995), the use of exemplars (Radnor, 1995) and the development of 'a community of practice' (Hall and Harding, 2002). Methods that depend on teachers meeting to share their assessment of specific pieces of work require time and resources, which were reported in many studies as no longer being provided for supporting TA, even though this had been the case in the past. Radnor (1995) noted that secondary teachers recognised the importance of teachers moderating each others' assessment, but this was difficult even among the teachers in the same school when priorities were directing attention and resources elsewhere. Thus teachers were working individually and using standard exemplar materials to guide their decisions.

Hall and Harding (2002) reported contrasting approaches among the six primary schools in their study. In some schools, time was allocated for teachers to meet and to work as a whole school in developing more accurate methods of assessing children. In these schools, there was, for example, recognition of the use of portfolios to communicate among teachers, with children and with parents about how work was assessed. In other schools, a decline in the level of collaboration was noted after an initial period in which time was made available to discuss assessment. Hall and Harding (2002) also noted that schools remained isolated from each other and also largely from the LEA assessment advisers. These advisers had developed considerable knowledge and expertise in TA, but found limited opportunities to share this with teachers. The advisers potentially had a key role in moderating TA work and forming a pathway through which teachers could share their understanding of assessment criteria and how to apply them to students' work. A resource which could have been used for moderation was left unused. Schools need to develop routines which make provision for assessment meetings within and across schools, since 'the

quality of teaching and learning inside the classroom is strongly influenced by the quality of professional relationships teachers have with their colleagues outside the classroom (Hargreaves and Evans, 1997; Anderson and Herr, 1999)' (Hall and Harding, 2002, p 12).

Main points relating to the conditions that affect the reliability and validity of teachers' summative assessment

Each of these points is supported both by evidence of high weight and evidence of medium weight.

- Several studies report bias in TA relating to student characteristics, including behaviour (for young children), gender, special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills (Bennett *et al.*, 1993; Reeves *et al.,* 2001; Thomas *et al.*, 1998; Shorrocks *et al.*, 1993; Brown *et al.*, 1996, 1998; Delap, 1994, 1995; Wilson and Wright, 1993; Levine *et al.,* 1987).

- There is variation in the level of TA and in the difference between TA and standard tests or tasks that is related to the school. The evidence is conflicting as to whether this is increasing or decreasing over time. There are differences among schools and teachers in approaches to conducting TA (Reeves *et al.,* 2001; Thomas *et al.*, 1998; Gipps *et al.*, 1996; Hall *et al.*, 1997; Hall and Harding, 2002).

- Evidence in relation to the reliability and validity of TA in different subjects is mixed. Differences between subjects in how TA compares with standard tasks or examinations results have been found, but there is no consistent pattern suggesting that assessment in one subject is more or less reliable than in another (Reeves *et al.,* 2001; Shorrocks *et al.*, 1993; Radnor, 1995; Delap, 1994, 1995; Levine *et al.,* 1987).

- It is important for teachers to follow agreed procedures if TA is to be sufficiently dependable to serve summative purposes. To increase reliability, there is a tension between closer specification of the task and of the conditions under which it is carried out, and the closer specification of the criteria for judging performance (Gipps *et al.*, 1996; Hall *et al.*, 1997, Hall and Harding, 2002; Radnor, 1995; Koretz *et al.*, 1994; Shapley and Bush, 1999; Rowe and Hill, 1996).

- The training required for teachers to improve the reliability of their assessment should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of the language used. Training should also focus on the sources of potential bias that have been revealed by research (Koretz *et al.*, 1994; Shapley and Bush, 1999; Levine *et al.*, 1987; Frederiksen and White, 2004; Rowe and Hill, 1996; Hargreaves *et al.,* 1996).

- Teachers can predict with some accuracy their students' success on specific test items and on examinations (for 16 year-olds) given specimen questions. There is less accuracy in predicting 'A' level grades (for 18 year-olds) (Coladarci, 1986;

Wilson and Wright, 1993; Delap, 1994, 1995; Good and Cresswell, 1988; Koretz *et al.*, 1994; Shapley and Bush, 1999).

- Detailed criteria describing levels of progress in various aspects of achievement enable teachers to assess students reliably on the basis of regular classroom work (Rowe and Hill, 1996; Meisels *et al.,* 2001; Frederiksen and White, 2004).

- Moderation through professional collaboration is of benefit to teaching and learning as well as to assessment. Reliable assessment needs designated time for teachers to meet and to take advantage of the support that others, including assessment advisers, can give. (Good, 1988; Radnor, 1995; Hall and Harding, 2002).

4.4 In-depth review: quality assurance results

Data-extraction for all 30 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, a more complete description was produced by combining aspects of detail provided by each extraction. In only two cases were there difference in study type, which had consequences for subsequent questions in the EPPI-Reviewer. Differences in judgements of weight of evidence occurred in four cases. In one, the judgement was mistakenly made on the quality of the assessment procedures being studied rather than the methodological quality of the study. In two cases, superior knowledge of statistics of one of the reviewers was brought to bear. In the fourth, one reviewer was unaware of alternatives that could have been adopted in the study to ensure greater trustworthiness. These examples strengthen the case for the quality assurance measures in the EPPI 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 teachers, professional developers, teacher educators and researchers. Despite the pressure on their time caused by the extra work, they reported unanimously that the 'hands-on' involvement was most valuable in understanding both the processes of the review, the nature of evidence available from research, and the issues surrounding the review question.

Users were able to have a wider involvement in the discussion of implications of the review for policy, practice and research. As reported in section 2.1.2, the findings of the review were discussed as part of an invitational seminar, held in Cambridge January 12th and 13th 2004. The 24 participants in the seminar included teachers and head teachers, researchers, representatives of teachers' organisations, of AAIA, and of UK government agencies involved in national assessment programmes. The findings of the review had been circulated in confidence in advance of the seminar, and after a brief overview discussion, groups were able to discuss the findings in

terms of their experience and to help to identify implications. Some points went beyond the review findings as participants linked the outcomes to their own experience, other research and specific events. In reporting implications, those directly arising from the review are listed separately from the additional points made at the seminar.

5. FINDINGS AND IMPLICATIONS

This chapter begins by summarising the outcomes of the review that have been presented in chapters 3 and 4. This is followed by a discussion of the findings from the in-depth review of studies with the aim of identifying key factors of the practice of using TA for summative assessment that are of relevance to users of the review. This discussion incorporates reference to relevant reviews of research and to other writing, which provides a useful background for the interpretation of findings. Some strengths and weaknesses of the current review are identified. Finally, some implications of the review findings are set out, drawing on consultation with a group of experiences practitioners, researchers and representative of government agencies.

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 studies found was 431, of which 369 were excluded in either a one-stage or a two-stage screening, using inclusion and exclusion criteria. Full texts were obtained for 48 of the remaining 62 studies, from which a further 15 were excluded during keywording, and two sets of studies (three in one case and two in the other) were linked as they were based on the same set of data. This left 30 studies after keywording. All of these were included in the in-depth review.

It was noted that while 79% of the original 431 studies were found through searching online databases, only 17% of the 30 included studies were from this source.

5.1.2 Mapping of all included studies

The 30 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; 15 were conducted in England, 12 in the United States and one each in Australia, Greece and Israel.

All studies were concerned with students between the ages of 4 and 18. Eleven involved primary school students (aged 10 or below) only, 13 involved secondary students (aged 11 or above) only, and six were concerned with both primary and secondary students. There was no variation across educational settings in relation to the focus of the study on reliability or validity, but there were slightly more evaluations of naturally-occurring situations in primary schools. Almost all studies in the primary and nursery school involved assessment of mathematics and a high proportion related to reading. At the secondary level, studies of assessment of

mathematics and 'other' subjects (variously concerned with foreign languages, history, geography, Latin and bible studies) predominated.

Eighteen 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 nine were either set externally or set by the teacher to external criteria. The majority were assessed by teachers using external criteria. The most common purpose of the assessment in the studies was for national or state-wide assessment programmes, with six studies relating to certification and another six relating to to informing parents (in combination with other purposes). A large proportion of the studies were concerned with national or state-assessment programmes.

As might be expected in the context of summative assessment, most research related to the use of external criteria by teachers, with little research on student self-assessment or teachers using their own criteria.

5.1.3 Summary of main findings from studies in the in-depth review

Findings addressing the main research question (what is the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?) are brought together here. They have been combined into a single list so that evidence of reliability and validity can be considered together.

The main findings are as follows:

- There is high-weight evidence that reliability of portfolio assessment where tasks were not closely specified was low (Koretz *et al.,* 1994, Shapley and Bush, 1999). There is also tentative evidence that estimates of construct validity of portfolio assessment, derived from evidence of correlations of portfolios and tests, were low (Koretz *et al.*, 1994; Shapley and Bush, 1999).
- High-weight evidence indicates that finer specification of criteria, describing progressive levels of competency, has been shown to be capable of supporting reliable TA, while allowing evidence to be used from the full range of classroom work (Rowe and Hill, 1996). There is conflicting medium-weight evidence as to the relationship between teachers' ratings of students' achievement and standardised test score of the same achievement when the ratings are not based on specific criteria (Hopkins *et al.*, 1985; Sharpley and Edgar, 1986). However, other medium-weight evidence suggests that teachers' judgements guided by checklists and other materials in the Work Sampling System have high concurrent validity for assessment of Kg to Grade 3 students (Meisels *et al.*, 2001).
- Studies of the NCA for students aged 6 and 7 in England and Wales in the early 1990s gave high-weight evidence of considerable error and of bias in relation to different groups of students (Shorrocks *et al.*, 1993; Thomas *et al.,* 1998). However, there is medium-weight evidence that the interpretation of correlations of TA and standard task results for seven year-olds should take into account the variability in the administration of the standard tasks (Abbott *et al.*, 1994).
- Other high-weight evidence indicates that the introduction of TA as part of the national curriculum assessment initially had a beneficial effect on teachers'

planning and was integrated into teaching (Hall *et al.*, 1997). Medium-weight evidence suggests, however, that in the later 1990s there was a decline in earlier collaboration among teachers and sharing interpretations of criteria, as support for TA declined and the focus changed to other initiatives (Hall and Harding, 2002).

- Study of the NCA for 11 year-olds in England and Wales in the later 1990s shows that results of TA and standard tasks agree to an extent consistent with the recognition that they assess similar but not identical achievements (Reeves *et al.,* 2001). This is despite medium-weight evidence of variation of practice among teachers in their approaches to TA, type of information used and application of national criteria (Gipps *et al.*, 1996; Radnor, 1995).
- High-weight evidence shows that the clearer teachers are about the goals of students' work, the more consistently they apply assessment criteria (Hargreaves *et al.*, 1996) and that teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching (Coladarci, 1986). This is supported by medium-weight evidence that teachers who have participated in developing criteria are able to use them reliably in rating students' work (Hargreaves *et al.*, 1996; Frederiksen and White, 2004).
- When rating students' oral proficiency in a foreign language, teachers are consistently more lenient than moderators, but are able to place students in the same rank order as experienced examiners (Good, 1988a; Levine *et al.*, 1987).
- High-weight evidence shows that the potential for a science project as part of 'A' level examinations to assess skills different from those used in regular laboratory work was reduced when the project assessment was changed from external to internal by teachers (Brown, 1998). However, medium-weight evidence suggests that teachers are able to score hands-on science investigations and projects with high reliability, using detailed scoring criteria (Frederiksen and White, 2004; Shavelson *et al.* 1992). Other medium-weight evidence shows that TA of practical skills in science makes a valid contribution to assessment at 'A' level within each science subject but there is little evidence of generalisability of skills across subjects (Brown *et al.*, 1996).
- There is medium-weight evidence that the rate at which young children can read aloud is a valid, curriculum-based measure of reading progress as measured by a standardised reading test (Crawford *et al.*, 2001).
- Medium-weight evidence suggests that teachers' perceptions of students' ability and probability of success on a test are moderately valid predictors of performance on the test, as are student self-assessments of their performance on a test after they have taken it (Wilson and Wright, 1993).

Findings in respect of the subsidiary review question (*What conditions affect the reliability and validity of teachers' summative assessment?*) were summarised in section 4.3. Each of these points is supported both by evidence of high weight and evidence of medium weight.

1. Several studies report bias in TA relating to student characteristics, including behaviour (for young children), gender, special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills (Bennett *et al.*, 1993; Reeves *et al.,* 2001; Thomas *et al.*, 1998; Shorrocks *et al.*, 1993; Brown *et al.*, 1996, Brown, 1998; Delap, 1994, 1995; Wilson and Wright, 1993; Levine *et al.,* 1987).

2. There is variation in the level of TA and in the difference between TA and standard tests or tasks that is related to the school. The evidence is conflicting as to whether this is increasing or decreasing over time. There are differences among schools and teachers in approaches to conducting TA (Reeves *et al.,* 2001; Thomas *et al.*, 1998; Gipps *et al.*, 1996; Hall *et al.*, 1997; Hall and Harding, 2002).

3. Evidence in relation to the reliability and validity of TA in different subjects is mixed. Differences between subjects in how TA compares with standard tasks or examinations results have been found, but there is no consistent pattern suggesting that assessment in one subject is more or less reliable than in another (Reeves *et al.,* 2001; Shorrocks *et al.*, 1993; Radnor, 1995; Delap, 1994, 1995; Levine *et al.,* 1987).

4. It is important for teachers to follow agreed procedures if TA is to be sufficiently dependable to serve summative purposes. To increase reliability, there is a tension between closer specification of the task and of the conditions under which it is carried out, and the closer specification of the criteria for judging performance (Gipps *et al.*, 1996; Hall *et al.*, 1997, Hall and Harding, 2002; Radnor, 1995; Koretz *et al.*, 1994; Shapley and Bush, 1999; Rowe and Hill, 1996).

5. The training required for teachers to improve the reliability of their assessment should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of the language used. Training should also focus on the sources of potential bias that have been revealed by research (Koretz *et al.*, 1994; Shapley and Bush, 1999; Levine *et al.*, 1987; Frederiksen and White, 2004; Rowe and Hill, 1996; Hargreaves *et al.,* 1996).

6. Teachers can predict with some accuracy their students' success on specific test items and on examinations (for 16 year-olds) given specimen questions. There is less accuracy in predicting 'A' level grades (for 18 year-olds) (Coladarci, 1986; Wilson and Wright, 1993; Delap, 1994, 1995; Good and Cresswell, 1988; Koretz *et al.*, 1994; Shapley and Bush, 1999).

7. Detailed criteria describing levels of progress in various aspects of achievement enable teachers to assess students reliably on the basis of regular classroom work (Rowe and Hill, 1996; Meisels *et al.,* 2001; Frederiksen and White, 2004).

8. Moderation through professional collaboration is of benefit to teaching and learning as well as to assessment. Reliable assessment needs designated time for teachers to meet and to take advantage of the support that others including assessment advisers can give (Good, 1988; Radnor, 1995; Hall and Harding, 2002).

5.2 Discussion of findings from studies in the in-depth review

5.2.1 Requirements for dependable summative assessment by teachers

Assessment in the context of education is a process of deciding, collecting and reasoning from evidence about learners' knowledge and skills. All assessment is based on a view of learning, a sample of behaviour in the domain of interest and a way of interpreting the behaviour in the domain of interest (NRC, 2001). Expanding this a little, it can be argued that for dependable summative assessment (that is, with construct validity protected and optimum reliability), the requirements are as follows:

- decisions about the domain of knowledge, skills and other attributes of learning to be assessed that are justified in terms of how learning takes place;
- a valid sample of student behaviour in the domain;
- criteria for judging the sample that are well matched to the goals of the work, of the curriculum and of the domain;
- procedures for the reliable and unbiased application of the criteria;
- procedures for reporting and communicating with users of the assessment outcomes.

The various approaches to using assessment by teachers for summative purposes represented in the studies in this review need to be considered in relation to how well they meet these requirements. First it is relevant to make two points. One is to recall that the meaning of assessment by teachers (TA) used here is that the teacher is involved both in gathering the evidence of achievement *and* in applying criteria to judge it. This excludes situations where work products from the classroom are sent for external marking or where teachers other than the students' own mark work, except in the context of moderating teachers' judgements. The second point is that the concern here is with TA only, not with complete systems of assessment of which TA may be a part. Thus, while some authors (e.g. Delap, 1995) have suggested that, for dependable assessment of achievement, TA should be supplemented by other measures such as standard tests, the concern here is only with the dependability of the TA component in such a system.

Some brief points of comment follow concerning the requirements before considering how they are met in the studies reviewed.

Decisions about what to assess based on a view of learning

The view of what it is to be a successful learner varies with how the process of learning is envisaged. Tests which require only knowledge that can be stored in short-term memory are implicitly based on a narrow view of learning which does not take account of current understanding of how knowledge and skills are built up, and particularly of how short-term and long-term memory interact with each other. Recognising the value of how learners organise their learning, rather than merely what they can recall means that assessment should 'address whether students know when, where and how to use their knowledge' (NRC, 2001, p 73). The structured tasks that are generally used in tests, and often in assessment by teachers, too, may

not reflect the way of learning nor give opportunity for students to show the skills and knowledge, that teachers value. Consider, for example, the case of assessment in the arts, reported by Hargeaves *et al.* (1996), in which there was a high level of internal consistency for structured activities but the products of unstructured activities were rated more highly than the products of structure activities of the same students.

Brookhart (1994, p 284) in her review of research on teachers' grading practices reports the in-depth case study by Briscoe (1991) showing how a teacher's beliefs informed his grading practices: 'Beliefs on which this teacher based his grading decisions included the following: (a) students should have opportunities to excel, (b) the school is a workplace, (c) individual students should be accountable, and (d) good teaching results in a low failure rate'.

A valid sample of behaviour

An assessment can only sample behaviour, yet it leads to statements that refer to the full range of knowledge and skills represented in the domain. Various types of validity were discussed in section 1.2, where it was argued that construct validity can be regarded as subsuming other types, and thus is paramount in determining the dependability (also see section 1.2) of assessment by teachers, just as it is for assessment carried out in other ways. So, if an assessment purports to measure, for instance, skills of scientific enquiry, there should be evidence that students are using such skills in the tasks assessed. In this context, it is important to distinguish between the opportunity to use skills and actually using them (Hamilton *et al.*, 1997). Thus analysis of the task demand (content validity) may not be enough. Other evidence might come from interviewing students to have them talk through their reasoning in working on a task, or from correlation with other measures or evidence of related learning (predictive or concurrent validity). It has also be suggested that assessments should be judged in relation to the value they have in making useful decisions (Messick, 1989). All these forms of evidence are brought together under the title of construct validity.

Many of the studies reviewed here showed that standard tests are often judged by teachers to provide too narrow a view of knowledge and skills to support the interpretations made of the results. At the same time, there is evidence that, if teachers make the selection of the sample, this, too, may not reflect the full extent of the domain. Stables (1992), who researched the assessment by teachers of speaking and listening as part of the assessment of English in the National Curriculum assessment of 14 year-olds in England, points out the need for agreement on what constitutes evidence in these aspects of achievement. The same may well be true of other aspects where agreement on evidence is assumed, rather than questioned.

Criteria that are well matched to the goals

In tests, it is taken for granted that marking schemes (or protocols) will match specific items. However, when assessment is not based on specific tasks or items, and instead a variety of tasks may provide the content in which the knowledge or skills to be assessed are shown, the relationship of the criteria to the evidence is more problematic and needs to be made explicit. Assessment criteria can have a dual function in assessment where evidence is, or can be, taken from a range of activities. One function is to focus attention on relevant evidence; the other is as a basis for interpreting and making judgements of the evidence in terms of the extent to which

the criteria are met. For valid assessment, it would seem obvious that it is important for the criteria to match the learning goals. This is not necessarily the same as matching the assessment tasks, unless the tasks are themselves an adequate sample of the full range of goals.

Procedures for arriving at reliable and unbiased judgements

To be dependable, the sample of behaviour assessed must provide adequate evidence to support the interpretations and judgements based on it. In the case of using TA for summative purposes, this means that both validity and reliability have to be optimised. All assessment is subject to error, which can be random or systematic. Random errors in assessment by teachers can have several causes relating to the identification of evidence, the understanding of criteria and the application of criteria. The evidence from studies reviewed suggests that teachers are more reliable in their assessment when they have a good grasp of the criteria, which will help identify relevant evidence as well as to make judgements of it (Hargreaves *et al.*, 1996; Rowe and Hill, 1996; Frederiksen and White, 2004).

Bias is a non-random source of error. In tests, this can arise on account of the form in which questions are put and the form in which answers are required. For example gender differences have been reported in relation to open-ended and multiple-choice item forms and in relation to the contextualisation of presented problems (Gipps and Murphy, 1994; Murphy, 1988 and 1993; Parker and Rennie, 1998). In assessment by teachers, there may be less bias due to unfamiliar situations, particularly if the assessment is embedded in regular work, but there is more opportunity for knowledge of non-relevant factors, such as behaviour, as well as gender and general performance, unconsciously to influence teachers' judgements.

For assessment by teachers, given that the range of regular activities provides equal opportunities for all to learn and to use and develop their knowledge and skills, ideally bias should be eliminated at the point of applying criteria. Efforts to do this include training in careful application of criteria to identify valid evidence and to makie judgements (Gipps and Murphy, 1994). Bias which exists after judgements have been made can be detected and controlled by moderation and adjustment of the judgements. This can be through comparison with judgements of the same evidence of others, particularly those who have been trained to avoid bias and have experience in looking across a number of teachers' judgements.

Procedures for reporting to, and communicating with, users of the assessment

Summative assessment, as opposed to formative assessment, is carried out to provide information to others than the teacher and the students; formative assessment, on the other hand, is for use mainly by the teacher and students involved. Thus attention needs to be given to reporting as well as to the processes of information gathering and judging. Cizek *et al.* (1995/6, p 161), reporting on teachers' grading practices, point out that even though assessment practices may have changed over the years, their achievement at the end of a period of time is still reported in terms of grades or marks. 'The largely unaddressed problem is that teachers' practices for assigning grades vary widely and unpredictably. The meaning of a student's grade to any interested party – the parents, other teachers, college admissions departments, employers, and even the student – is unclear'. In their

research, Cizek *et al.* (1995/6) found that only about 50% of teachers surveyed were aware of the grading policies which they were expected to follow.

Similar problems of communication occur where there are stated criteria and procedures for the teacher to follow. In reporting in the context of the National Curriculum, for instance, level labels are used to represent the criteria used in assessing achievement. The meaning of these may not be known to the recipients of the information and are thus interpreted by them in varying and unpredictably ways. Also, as Shapley and Bush (1999) found, many teachers do not follow intended procedures even when these are set out.

5.2.2 Variation in procedures for assessment by teachers and their impact on dependability

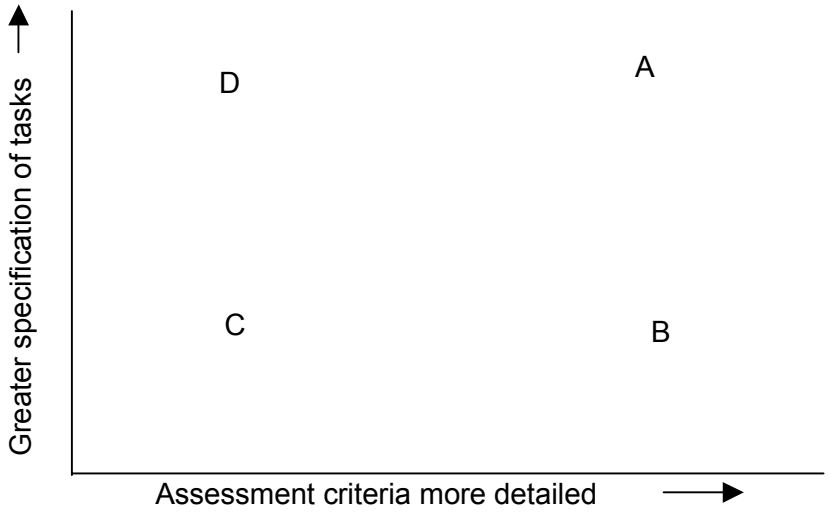
The procedures for assessment by teachers studied in the research included in this review illustrate the range of approaches that have been or are being practised. They also give some indication of whether, and if so how, the differences affect dependability (that is, the interconnected constructs of reliability and validity). The major differences are in:

- how the assessment is set up (through the degree of specification of the student work used as evidence and the detail in which the criteria used in making judgements are specified)
- the way in which teachers and schools use the procedures.

Dimensions of variation in how the procedures are set up

The variation in the specification of the task or tasks can be envisaged as spread along a dimension from unspecified to highly specified. For each type of task, theoretically, there are different approaches spread along another dimension, from loosely specified criteria to closely specified criteria for judgement. Figure 5.1 indicates four main types of approaches that are defined by the intersection of these dimensions.

Figure 5.1: Classification of approaches to summative assessment defined by degree of tasks specification and detail of criteria



The vertical 'task' dimension (y axis) extends from no specification, when the assessment is based on the whole range of regular work to tight specification, as in tests, passing through a mid-point where types of tasks to be included in the assessment may be specified. The 'criterion' dimension (x axis) extends from general judgements, grading or ratings where no precise meaning is given to the labels used, to detailed criteria which match particular tasks, passing through a mid-point where brief descriptions are used to define points on a grading or rating scale.

- The approach in area A combines a high degree of specificity of the task with high detailed criteria. Such approaches are, or are close to being, tests, where marking criteria are matched to specific items. Examples are oral tests of performance in a foreign language or practical science investigations which are specified externally but assessed by teachers (Good, 1988a; Shavelson *et al.*, 1992).

- In area B, detailed criteria are combined with evidence from a range of different kinds of task, which could be found in regular work tasks. Examples are the use of detailed developmental criteria in relation to regular work, such as in the Australian 'Developmental Assessment' (Rowe and Hill, 1996).

- In area C, there is little close specification of either criteria or tasks, leaving teachers to select work to be assessed and to relate broadly stated criteria in judging it. Examples are the National Curriculum Assessment in England, using level descriptions, as described in Hall and Harding (2002).

- In area D, tasks are defined, while criteria are in general terms. In practice these are uncommon, since the specification of tasks is usually accompanied by specific criteria.

When tasks are unspecified, tight criteria can guide the selection of work assessed (area B), while general, non-specific criteria leave the validity of the sample in the hands of the teacher (C). When tasks are specified (D and A), the validity depends on the selection made in designing the assessment programme and on how well the criteria match the specified tasks.

The degree to which tasks for assessment by teachers can, or should be specified, goes to the heart of reasons for including TA in assessment systems rather than depending on standard tests and tasks which are externally marked. It is recognised that what can be assessed by teachers is different from what is assessed by tasks designed to be applied uniformly to all pupils (Reeves *et al.*, 2001; Sharpley and Edgar, 1986). Therefore this discussion refers to the extent of task specification. The variations relating to specification of criteria are considered within groups identified by the degree of specification of the tasks (that is, cutting across Figure 5.1 horizontally).

Studies where the tasks are specified (A and D)

An issue raised by Abbott *et al.* (1994) in relation to the assessment of young children is the extent to which it is realistic to specify tasks so that all children have comparable opportunities to demonstrate their knowledge and skills. Although raised

in the context of teachers administering standard performance tasks, the issue is relevant in all cases there teachers both set up the assessment tasks and collect evidence from them. In practice, unless the class is turned into an examination room, thus nullifying the advantages of normal work that assessment by teachers allows, conditions can be far from uniform across all pupils, even if the tasks are tightly specified.

Assessment approaches using portfolios where there is a requirement to include examples of certain kinds of work fall into this category of specified tasks. However, in the portfolio systems approach initially tried in Vermont (Koretz *et al.*, 1994) and in Texas (Shapley and Bush, 1999), teachers were given a relatively free choice of what to include and asked to rate the work in terms of how far certain goals were achieved. This placed these approaches in area C of Figure 5.1. The reported reliability of the judgements made using these approaches was very low. Steps taken to tighten the guidelines for selecting work and for scoring it, moving it into area A, were not successful in raising the reliability to a level where the measures could be used for reporting individual achievement, or for reporting aggregate scores across groups. The low reliability was ascribed to the lack of match between the tasks and the criteria (which were in general terms) and the inconsistency in teachers' application of the criteria. The measures needed to improve the situation, suggested by Shapley and Bush (1999), were to prescribe more closely the work samples so that matching criteria could be used. They suggested that a 'core' set of tasks should be included, thus moving the approach to area A. This carries the risk of attention being focused on these pieces of work, just as it can be on what is tested by external tests, especially when the outcome is used for a purpose that has high stakes for the teachers.

In the Vermont portfolio programme, as Koretz *et al.* (1994) reported, the solution to the problem of low reliability was to have scoring carried out by teachers other than the students' own, who were trained in applying the criteria, given that training would be more efficient with scorers gathered together. (This takes the Vermont portfolio programme outside the definition of TA used in this review.) However, although the inter-rater reliability improved over time, it remained low and this was ascribed to the difficulty of training a large number of scorers and the lack of standardisation of tasks. A further difficulty reported by Shapley and Bush (1999), when checking on teachers' scoring, was that the necessary information about the context and goals of the pieces of work in the portfolios was often not provided. Teachers were not adhering to the requirements for selecting and labelling work in the portfolios.

Evidence that teachers can use criteria consistently when these are designed for specific types of performance comes from studies in the visual arts, music and science projects (Hargreaves *et al.*, 1996; Frederiksen and White, 2004; Shavelson *et al.*, 1992). Hargreaves *et al.* (1996) also provide high-weight evidence that the more thoroughly teachers understand the criteria the more consistently they apply them. However, Frederiksen and White (2004) found that prior experience of scoring with criteria is not a prerequisite for consistent use (evidence of medium weight). This suggests that an essential difference between the Vermont and Texas experiences and those of Hargreaves *et al.* (1996), Frederiksen and White (2004) and Shavelson *et al.* (1992) lies not in whether or not training is given, but in the type of training given to the teachers. This point is reinforced by the experience of Gilmore (2002) referred to in section 1.3.2. Although this was in the context of training test

administrators and markers, Gilmore involved administrators in working with children and markers in developing a mark scheme. As a result, greater understanding was developed than through training in how to use predetermined criteria consistently.

Studies where tasks are not specified (areas B and C of Figure 5.1)

The subject profile approach studied by Rowe and Hill (1996) allows work to be gathered from the full range of classroom work and provides detailed criteria for judging achievement in various subjects and strands within subjects. Each set of criteria describes development in relation to a particular type of achievement. Teachers are not trying to match a particular piece of work with a particular criterion, but taking evidence from several relevant pieces and forming a judgement on the basis of the best match between the evidence and the criteria. The approach falls in the area B in Figure 5.1. The criteria, however, serve the additional function of focusing attention on the outcomes of particular kinds of work, so that teachers are alerted to looking for particular behaviours and are less likely to miss them. Rowe and Hill (1996) report consistency across judgements made by the same teachers on a different occasion.

Evidence from other studies (Meisels *et al.,* 2001; Coladarci, 1986) also suggests that when criteria are well specified, teachers are able to make reliable judgements. This is reinforced by the evidence that, when the criteria are more general, and so are not matched to particular pieces of work (that is, area D), reliability is low (Koretz *et al.;* 1994; Shapley and Bush, 1999). Further, in the approach studied by Sharpley and Edgar (1986), the indication is that the judgements made using general criteria have low concurrent validity, being poorly aligned with other assessment of the same achievements.

The assessment by teachers (TA) in the National Curriculum Assessment in England and Wales allows evidence to be used from the regular classroom work. There is evidence that how teachers go about this varies (Hall *et al.*, 1997; Gipps *et al.*, 1996; Radnor, 1995), but this does not in itself necessarily affect the reliability. Teachers vary in their teaching approaches and any less variation in assessment practice would not be expected. Certainly, variation according to the nature of the subject and how it is taught, as noted by Radnor (1995), is to be expected if assessment is truly embedded in regular work.

Brookhart's (1994) conclusions from her review of research in grading practices support these findings. She found considerable variation among teachers, particularly in relation to taking aspects such as effort into account. There was also a difference between elementary teachers, who used more informal evidence and observation, and secondary teachers who used more paper-and-pencil achievement measures. Frary *et al.* (1993) surveyed the grading practices of a random sample of secondary teachers in Virginia, USA. They found that teachers based their grades on evidence from a range of sources, including teacher-made tests, quizzes, judgements of overall ability, and effort as well as ongoing grades for class work. They also used cluster analysis of teachers' to identify groups of teachers, as did Gipps *et al.* (1996). The groups found by Frary *et al.* (1993) were described as norm referencing, softhearted, strict, arbitrary, uncertain and inconsistent. Commenting on this, Brookhart (1994, p 285) calls for more research into the reasons for grading in certain ways and attention to the links between 'assessment recommendations and the broader issues and purposes of learning and instruction'.

The criteria used by teachers in the National Curriculum Assessment in England has, since 1995, been in the form of 'level descriptions' for each subject sub-division, which are not unlike less detailed versions of the developmental criteria described by Rowe and Hill (1996) and which, according to Reeves *et al.* (2001), are used with consistency, at least by teachers of 11 year-olds, and provide a useful complement to standard tests. Since these descriptions are criteria that can be applied to a range of events in regular work, the approach can be described as being on the border between areas C and B in Figure 5.1. However, there is often uncertainty as how the evidence is gathered and, when teachers depend on evidence from classroom tests in order to judge achievement, the approach moves to the border of areas D and A and even into area A. Then TA becomes in effect a series of standard tasks created by the teacher, or based on imported tests, which may not reflect either the full range of the curriculum nor what pupils can do when not under test conditions. There is a danger compromising construct validity by restricting the range of evidence that is taken into account. Moreover, a study of teacher-made tests in mathematics and science by McMorris and Boothroyd (1993) found them to be of low quality, with flaws in 35% of completion items and in 20% of multiple-choice items. They found that the quality of teachers' tests was related to teachers' competence in measurement and that teachers who had attended measurement courses produced better tests, a finding confirmed by Plake *et al.* (1992).

The validity of approaches which leave unspecified the sampling of the domain depends on the extent to which the evidence actually gathered is a good sample of work in the areas concerned. While having specific criteria leads teachers to consider work from relevant areas, if these are not well covered in the implemented curriculum, the opportunity for assessment is clearly limited. There is evidence that consistency in applying criteria seems to depend upon teachers being clear about the goals of the work (Hargreaves *et al.,* 1996; Koretz *et al.,* 1994) and on the thoroughness with which relevant areas of the curriculum are covered in teaching (Coladarci, 1986). The potential for consistent use of criteria, as found by Reeves *et al.* (2001), does not mean that the criteria will be used in relation to an adequate sample of the domain. There is also a further aspect to the selection of tasks, pointed out by Frederiksen and White (2004), that, for valid assessment, the tasks need to be engaging and meaningful to the pupils.

Hall and Harding (2002) identified the context of the school's support and value system as having a role in how TA is practised. Conditions having relevance for the validity of the assessment include the extent to which teachers share interpretations of criteria and develop a common language for describing and assessing pupils' work.

The dual role of teachers: teacher or assessor

A feature of the classroom in which teachers conduct assessment for summative purposes is the possible conflict in roles that this creates. This was mentioned in the Background to this review (section 1.3), but, perhaps surprisingly, has not featured to any large extent in the studies reviewed in depth. Choi (1999) noted that this was likely to be a problem in systems where the summative assessment has 'high stakes' for the students. The issue was also raised by Stables (1992) in the context of National Curriculum assessment at Key Stage 3.

Morgan (1996) studied the approaches of secondary mathematics teachers to the task of assessing the coursework of students which was assessed by teachers as part of the GCSE examination. Although this study did not report on validity and reliability issues, it showed that teachers go about the task of assessing specific pieces of coursework in quite different ways. One approach is to apply the assessment criteria strictly, taking on the role of the examiner and keeping to the evidence. Another is to attempt to understand what the student was trying to do in the work, engaging with the problem the student was tackling and recognising where the student 'might improve his work rather than merely pointing to its limitations' (p 365). Teachers who take the latter approach tend to hold to the role of the teacher, forming a picture of a student's ability and awarding the grade to the student rather than to the particular piece of work. This might be called a 'teacherly' approach, as opposed to an 'assessorly' approach. It has some relevance to the use of 'level descriptions' in the National Curriculum Assessment studied by Hall and Harding (2002). In using these descriptions, teachers look across work (and therefore at the students) not at individual pieces of work – a teacherly rather than an assessorly approach.

Morgan points out that adopting a strict examiner role may well conflict with teachers' value systems and understanding of how learning takes place, which is relevant to the reliability and validity of summative assessment. She quotes Galbraith's (1993) comment that 'the 'constructivist' paradigm associated with the current discourse of mathematics teaching is not compatible with the 'conventional' paradigm of external examinations'.(p 368) However, Morgan found that, in most of the cases she studied, the different basis for judgement did not have major effects on the outcomes of the assessment. She questions, however, the meaning that can be applied to the outcome in relation to the students' mathematical attainment.

5.2.3 Bias in assessment by teachers

Evidence of bias in TA comes mainly from studies where TA is compared with another measure and based on the questionable assumption that the benchmark measure is unbiased and is measuring the same thing as the TA. So, while it has been reported that teachers under-rate boys more than girls in mathematics and science, compared with their performance in tests (Reeves *et al.,* 2001), the conclusion might equally be that boys perform above expectation on mathematics and science tests. This could be, for instance, due to boys having better test-taking skills in these areas. Similarly, several studies report TAs of students with special educational needs (SEN) being below their score levels on tests of the same achievements (Reeves *et al.*, 2001; Shorrocks *et al.*, 1993; Thomas *et al.*, 1998). On the assumption that standard tasks are unbiased, there is evidence that TA varies systematically, but the same differences found in relation to gender, first language and SEN have been found by the same authors in the results of standards tasks.

Using a different approach, of testing a theoretical model of the relationship between TA and student variables and using path analysis to identify the effects on teachers' judgements, Bennett *et al.* (1993) found that teachers' view of young students' behaviour had a consistent effect. This disadvantaged boys, whose behaviour was judged to be poorer than that of girls, and led to a lower judgement of academic performance. It is possible that this impact is particular to young children and indeed Bennett *et al.* (1993) found a smaller effect in second grade than in first grade.

For students at the other end of the age range considered here, Brown (1998) and Brown *et al.* (1996) found some evidence of a 'halo' effect in teachers' judgements of specific skills used in science investigations. In one case, the assessment of skills made by teachers during regular laboratory work appeared to influence the assessment of skills used in a specific project (Brown, 1998). In Brown *et al.* (1996), the teachers' judgements of the students' overall ability appeared to influence judgements of particular skills.

In the assessment of oral proficiency in a foreign language, there is strong evidence that teachers over-rate their students, compared with external moderators, but that they do this consistently, placing the students in the same rank order as experienced examiners (Good, 1988a; Levine *et al.* 1987). Levine *et al.* (1987) also found systematic greater overestimation for higher achieving pupils, compared with those in the middle and lower achieving levels. These authors also make explicit what is implied by several other researchers, that training in closer and more conscious use of criteria is needed to reduce bias. The experience of Good (1988a) and Brown (1998) would suggest that moderation is also an important requirement and Radnor (1995) found that teachers favoured it.

Hoge and Butcher (1984) criticise approaches to validating teachers' judgements by correlation with tests, since different domains may be being assessed. For this reason, in their study, as in that of Coladarci (1986), teachers were asked to provide direct estimates of the performance of students on the achievement test items. Greater accuracy was found for higher achieving students than for lower achieving students. The authors concluded that the accuracy of teachers' judgements is influenced by teacher, student and task variables.

5.3 Strengths and weaknesses of this systematic review

5.3.1 Strengths

The review procedures enabled studies to be selected according to strict inclusion and exclusion criteria, based on the focus and content of the research rather than the design or types of evidence collected. This resulted in a range of studies of different designs, bringing different kinds of evidence to bear on the question posed for the review. Thus, some studies provided qualitative evidence relevant to the validity of the assessment carried out by teachers, while others provided quantitative evidence of the relationship between the TA and other measures of students' achievement. Since both types of evidence are needed to address the review question, this is regarded as a strength of the review.

The quality assurance procedures of the review meant that all decisions, from the initial exclusion of studies from among those captured in the search to the final selection of studies for in-depth data-extraction, were checked through double independent action and documented. The application of inclusion and exclusion criteria, the application of keywords, the data-extraction and weight of evidence judgements were performed by at least two people working independently. Any differences in judgements were reconciled before findings were recorded and stored in the EPPI-Reviewer database. These measures enable the users of this review to

have confidence that relevant evidence has been reliably extracted at least from those studies identified in the initial search.

The studies selected for in-depth review cover a range of age of student, context, subjects assessed and purposes of assessment. This means that, while some generalisations are of limited extent, a number of relevant issues have been identified.

Theoretically, the evaluation of weight of evidence provided by each study is a strength of the review, since the review findings are not 'diluted' by evidence that is either somewhat suspect or drawn from situations not entirely relevant. However the difficulty of making judgements and particularly of combining them into one overall judgement makes the process somewhat problematic.

5.3.2 Limitations

No studies published before 1985 were included. Any cut-off is, of course, artificial and is particularly harsh when following up citations in already selected studies which draw attention to apparently relevant studies published just a year too early.

Attempts to conduct a comprehensive search for studies will inevitably fall short of the ideal. In terms of access to studies, US literature was accessed mainly through the ERIC and other databases, with fewer opportunities to handsearch journals. Since the latter provides a greater proportion of the studies eventually included in the review (see section 5.1.1), there is a doubt that some articles not listed in ERIC may have been overlooked. However, the large proportion (50%) of studies from the UK is probably explained by the interest in the UK in the focus of this review since the introduction of assessment by teachers as part of national assessment programmes and the consequently high number of research studies which have been focused on it.

Although some relevant studies will have been missed, and 14 out of the 62 papers of potential relevance could not be obtained, the range and findings, which by no means all point in one direction, support the view that a range of different experiences and conclusions has been tapped.

5.4 Implications

In order to explore the implications of the review findings for policy, practice and research, the synthesis of outcomes was presented at a seminar attended by expert practitioners, researchers and representatives of agencies concerned with policy in assessment. The event was held in Cambridge on January 12th and 13th 2004. Of the 24 participants in the seminar, eight were the members of the Assessment Reform Group, four were teachers or head teachers, two were representatives of teachers' organisations, six were from UK government agencies involved in national assessment programmes, one was the vice-president of the AAIA, one from an examinations board, one a university researcher and one from the Nuffield Foundation, which funded the project of which this seminar was one event. A detailed account of the findings of the review was circulated in confidence in advance of the seminar, giving participants the opportunity to study the review outcomes in

depth in preparation for discussion of their implications during a two and a half hour session.

One of the questions addressed in the group discussions following the presentation was 'How do the findings resonate with your experience?' The responses indicated that participants' experience did indeed concur with the findings of the review. The discussions brought out some points to be emphasised in drawing implications from the review findings. In some cases, the points went beyond the review findings in linking the outcomes to other experience, research and specific events. In the following sections, the implications that emerge directly from the review findings are set out first, followed by additional points raised in the seminar.

Solutions to the problems of inconsistency in the type of evidence used and in the application of criteria suggested by the studies focused on five types of action, relating to: the specification of the tasks, the specification of the criteria, training, moderation, and the development of an 'assessment community' within the school allied to increased confidence in the profession's judgement of teachers. These have implications for policy, practice and research, which are now set out.

5.4.1 Policy

(a) When deciding the method, or combination of methods, of assessment for summative assessment, the shortcomings of external examinations and national tests need to be borne in mind.
(b) The essential and important differences between TA and tests should be recognised by ceasing to judge TA in terms of how well it agrees with test scores.
(c) There is a need for resources to be put into identifying detailed criteria that are linked to learning goals, not specially devised assessment tasks. This will support teachers' understanding of the learning goals and may make it possible to equate the curriculum with assessment tasks.
(d) It is important to provide professional development for teachers in undertaking assessment for different purposes that address the known shortcomings of TA.
(e) The process of moderation should be seen as an important means of developing teachers' understanding of learning goals and related assessment criteria.

5.4.2 Practice

(a) Teachers should not judge the accuracy of their assessments by how far they correspond with test results but by how far they reflect the learning goals.
(b) There should be wider recognition that clarity about learning goals is needed for dependable assessment by teachers.
(c) Teachers should be made aware of the sources of bias in their assessments, including the 'halo' effect, and school assessment procedures should include steps that guard against such unfairness.
(d) Schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers is sustained (e.g. by protecting time for planning assessment, in-school moderation, etc.).

(e) School should develop an 'assessment culture' in which assessment is discussed constructively and positively and not seen as a necessary chore (or evil).

5.4.3 Research

(a) There should be more studies of how teachers go about assessment for different purposes, what evidence they use, how they interpret it, etc.
(b) The reasons for teachers' over-estimation of performance as compared with moderators' judgements of the same performance need to be investigated, to find out, for instance, whether a wider range of evidence is used by the students' own teachers, or whether criteria are differently interpreted.
(c) More needs to be known about how differences between schools influence the practice and dependability of individual teachers.
(d) Since evaluating TA by correlation with test results is based on the false premise that they assess the same things, other ways need to be found for evaluating the dependability of TA.
(e) There needs to be research into the effectiveness of different approaches to improving the dependability of TA, including moderation procedures.
(f) Research should bring together knowledge of curriculum planners, learning psychologists, assessment specialists and practitioners to produce more detailed criteria that can guide TA.

5.4.4 Additional points arising from consultation on the review findings with practitioners, policy-makers and researchers

(a) It is important to consider the purpose of assessment in deciding the strengths and weaknesses of using TA in a particular case. For instance, when assessment is fully under the control of the school and is used for informing pupils and parents of progress ('internal purposes), the need to combine TA with other evidence (e.g. tests) may be less than when the assessment results are used for 'external' purposes, such as accountability or the school or selection or certification of students.
(b) There needs to be greater recognition of the difference between purposes of summative assessment and of how to match the way it is conducted with its purpose. For instance, the 'internal' assessment that is under the control of the school should not emulate the 'external' assessment which has different purposes.
(c) If tests are used, they should be reported separately from TA, which should be independent of the test scores.
(d) There is evidence that a change in national assessment policy is due. The current system is not achieving its purpose. The recent report on comparability of national tests over time (Massey *et al.*, 2003) concludes that TA have shown less change in standards than the national tests. The authors state that 'National testing in its current form is expensive, primarily because of the external marking of the tests, and the time may soon come when it is thought that these resources may make a better contribution elsewhere' (Massey *et al.*, 2003, p 239).

(e) Improving teachers' formative assessment would also improve their summative assessment. Thus a programme of professional development aimed at enabling teachers' judgements to be used for summative purposes should be combined with attention to improving formative assessment.
(f) The role that pupils can take in their own summative assessment needs to be investigated and developed.
(g) Any change in the current systems requires a major switch in resources from test development to supporting teacher-led assessment.
(h) Change towards greater use of TA for summative purposes, requires a long-term strategy, with strong 'bottom-up' elements and provision for local transformations.

6. REFERENCES

6.1 Studies included in the map and data-extraction

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 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.

Brown CR (1998) An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England. *Studies in Educational Evaluation* **24:** 87-98.

Brown CR, Moor JL, Silkstone BE, Botton C (1996) The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations. *Assessment in Education* **3:** 377-391.

Chen M, Ehrenberg T (1993) Test scores, homework, aspirations and teachers' grades. *Studies in Educational Evaluation* **19:** 403-419.

Coladarci T (1986) Accuracy of teachers' judgements of students responses to standardised test items. *Journal of Educational Psychology* **78:** 141-146.

Crawford L, Tindal G, Steiber S (2001) Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment* **7:** 303-323.

Delap MR (1994) An investigation into the accuracy of A-level predicted grades. *Educational Research* **26:** 135-149.

Delap MR (1995) Teachers' estimates of candidates' performance in public examinations. *Assessment in Education* **2:** 75-92.

Frederiksen J, White B (2004), Designing assessment for instruction and accountability: an application of validity theory to assessing scientific inquiry. In: Wilson M (ed.) *Towards Coherence between Classroom Assessment and Accountability, 103rd Yearbook of the National Society for the Study of Education part II.* Chicago, IL, USA: National Society for the Study of Education.

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

Good FJ (1988a) Differences in marks awarded as a result of moderation: some findings from a teacher assessed oral examination in French. *Educational Review,* **40:** 319-331.

Good FJ, Cresswell M (1988) Can teachers enter candidates appropriately for examinations involving differentiated papers? *Educational Studies* **14:** 289-297.

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

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.

Hargreaves DJ, Galton MJ, Robinson S (1996) Teachers' assessments of primary children's classroom work in the creative arts. *Educational Research* **38:** 199-211.

Hopkins KD, George CA, Williams DD (1985) The concurrent validity of standardised achievement tests by content area asing teachers' ratings as criteria. *Journal of Educational Measurement* **22:** 177-182.

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.

Levine MG, Haus GJ, Cort D (1987) The accuracy of teacher judgement of the oral proficiency of high school foreign language students. *Foreign Language Annals* **20:** 45-50.

Meisels SJ, Bickel DD, Nicholson J, Xue Y, Atkins-Burnett S (2001) Trusting teachers' judgements: a validity study of a curriculum-embedded performance assessment in kindergarten to Grade 3. *American Educational Research Journal* **38:** 73-95.

Papas G, Psacharopoulos G (1993) Student selection for higher education: the relationship between internal and external marks. *Studies in Educational Evaluation* **19:** 397-402.

Radnor HA (1995) *Evaluation of Key Stage 3 Assessment Arrangments for 1995. Final Report.* Exeter: University of Exeter.

Reeves DJ, Boyle WF, Christie T (2001) The relationship between teacher assessment and pupil attainments in standard test/tasks at Key Stage 2, 1996-8. *British Educational Research Journal* **27:** 141-160.

Rowe KJ, Hill PW (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles'. *Assessment in Education* **3:** 309-352.

Shapley KS, Bush MJ (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience. *Applied Measurement in Education* **12:** 11-32.

Sharpley CF, Edgar E (1986) Teachers' ratings vs standardised tests: an empirical investigation of agreement between two indices of achievement. *Psychology in the Schools* **23:** 106-111.

Shavelson RJ, Baxter GP, Pine J (1992) Performance assessments: political rhetoric and measurement reality. *Educational Researcher* **21:** 22-27.

Shorrocks D, Daniels S, Staintone R, Ring, K (1993*) Testing and Assessing 6 and seven year-olds. The Evaluation of the 1992 Key Stage 1 National Curriculum Assessment*. UK: National Union of Teachers and Leeds University School of Education.

Thomas S, Madaus GF, Raczek AE, Smees R (1998) Comparing teacher assessment and the standard task results in England: the relationship between pupil characteristics and attainment. *Assessment in Education* **5:** 213-246.

Wilson J, Wright CR (1993) The predictive validity of student self-evaluations, teachers' assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachia schools. *Educational and Psychological Measurement* **53:** 259-270.

6.2 Other references used in the report

Anderson GL, Herr K (1999) The new paradigm wars: is there room for rigorous practitioner knowledge in schools and universities? *Educational Researcher* **28:** 12-21.

Black HD (1986) Assessment for learning. In: Nuttall DL (ed.) *Assessing Educational Achievement*. London: Falmer Press.

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.

Briscoe C (1991) Making the grade: multiple perspectives on a teacher's assessment practices. Paper presented at the annual meeting of the American Educational Research Association (AERA). Chicago, IL, USA: April 3-7.

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.

Butler R (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest. *British Journal of Educational Psychology* **58:** 1-14.

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

Cizek GJ, Fitzgerald SM, Rachor RE (1995/1996) Teachers' assessment practices: preparation, isolation, and the kitchen sink. *Educational Assessment* **3:** 159-179.

Crooks T, Flockton L (1993) Some proposals for national monitoring of education outcomes. University of Otago, Dunedin, New Zealand: Unpublished report.

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

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.

Evidence for Policy and Practice Information and Co-ordinating Centre (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.

Frary RB, Cross LH, Weber LJ (1993) Testing and grading practices and opinions of secondary teachers of academic subjects: implications for instruction in measurement. *Educational Measurement: Issues and Practice* **12:** 23-30.

Galbraith P (1993) Paradigms, problems and assessment: some ideological implications. In: Niss M (ed.) *Investigations Into Assessment in Mathematics Education: An ICMI Study*. Dordrecht, Netherlands: Kluwer Academic Publisher, pages 73-89.

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* **9:** 343-362.

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

Gipps C, Murphy PM (1994) *A Fair Test?* 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.

Griffin PE (1989) *Developing Literacy Profiles*. Coburg, Victoria, Australia: Assessment Research Centre, Phillip Institute of Technology.

Good F (1988b) A method of moderation of school-based assessments: some statistical considerations. *The Statistician* **37:** 33-49.

Hamilton LS, Nussbaum EM, Snow RE (1997) Interview procedures for validating science assessment. *Applied Measurement in Education* **10:** 181-200.

Hargreaves A, Evans R (1997) *Beyond Educational Reform: Bringing Teachers Back In.* Buckingham: Open University Press.

Harlen W (ed.) (1994) *Enhancing Quality in Assessment.* London: Paul Chapman.

Harlen W (1995) Standards and science education in Scottish schools. *Studies in Science Education* **26:** 107-134.

Harlen W (2000) *Teaching, Learning and Assessing Science 5 – 12.* London: Paul Chapman.

Harlen W, Deakin Crick R (2002) A systematic review of the impact of summative assessment and tests on students' motivation for learning. In: *Research Evidence in Education Library*. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Hoge RD, Butcher R (1984) Analysis of teacher judgements of pupil achievement levels. *Journal of Educational Psychology* **76:** 777-781.

Hoge RD, Coladarci T (1989) Teacher-based judgements of academic achievement: a review of literature. *Review of Educational Research* **59:** 297-313.

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

Johnson, S (1989) *National Assessment: The APU Science Approach*. London: Her Majesty's Stationery Office (HMSO).

Jones C, Strivens J (1991) Unit accreditation: a response to curriculum and assessment need. *Studies in Educational Evaluation* **17:** 275-289.

Klein SP, McCaffrey B, Koretz D (1995) The reliability of mathematics portfolio scores: lessons from the Vermont experience. *Applied Measurement in Education* **8:** 243-260.

Koretz D (1998) Large-scale portfolio assessment in the US: evidence pertaining to the quality of measurement. *Assessment in Education* **5:** 309-334.

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 AERA (Chicago, IL, USA: April 3-7) and the National Council on Measurement in Education (Chicago, IL, USA: April 4-6).

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**.

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 (QCA).

Masters GN, Forster MN (1995) *ARK Guides to Developmental Assessment.* Camberwell, Victoria: The Australian Council for Educational Research.

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

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.

Murphy PM (1988) Gender and assessment. *Curriculum* **9:** 165-171.

Murphy PM (1993) Some teacher dilemmas in practising authentic assessment. Paper presented to the AERA conference. Atlanta, GA, USA: April 12-16.

Murphy RL (1979) Teachers' assessments and GCE results compared. *Educational Research* **22:** 54-59.

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

Organisation for Economic Co-operation and Development (OECD) (1999) *Measuring Student Knowledge and Skills.* Paris: OECD.

OECD (2001) *Knowledge and Skills for Life: First Results from the PISA 2000*. Paris: OECD.

Parker LH, Rennie LJ (1998) Equitable assessment strategies. In: Fraser BJ, Tobin KG (eds.) *International Handbook of Science Education.* Dordrecht, Netherlands: Kluwer Academic Publishers, pages 897-910.

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, CA, USA: April.

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

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

Satterley D (1994) The quality of external assessment. In: Harlen W (ed.) *Enhancing Quality in Assessment*. London: Paul Chapman.

Spear MG (1984) Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education* **6:** 368-377.

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

Wiliam D (1993) Reconceptualising validity, dependability and reliability for National Curriculum assessment. Paper given at the British Educational Research Association (BERA) conference. Liverpool University: September.

Wiliam D (1995) It'll all end in tiers! *British Journal of Curriculum and Assessment* **5:** 21-24.

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

Yung B H-W (2002) Same assessment, difference practice: professional consciousness as a determinant of teachers' practice in a school-based assessment scheme. *Assessment in Education* **9:** 97-118.

APPENDIX 1.1: Review Group membership

The members of the Review Group and their affiliations are given below.

Members of Assessment Reform Group (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

International experts advising the Assessment and Learning Research Synthesis Group (ALRSG)

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

Ms Dina Kiwan, Education Research Officer (until September 2003)
Ms Zoe Garrett, Research Officer (from September 2003)
Rebecca Rees (from September 2003)

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 be 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
Studies were included if they reported information about the validity and/or reliability of methods used by teachers for summative assessment. 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 reliability or validity
B: not summative assessment
(aptitude and special needs assessment tests excluded; formative only excluded, but formative with summative included)
C: not teacher assessment
(assessment of teachers and school evaluation excluded)
D: not related to education in school
(higher education, nursing education, other vocational excluded)
E: not research
(instrument development excluded; also handbooks and reviews)

APPENDIX 2.2: Search strategy for electronic databases

Key terms

Combination of the terms listed below were used in searching ERIC and BEI from 1985 onwards.

For example, the search strategy sought studies identified by
{exp test reliability or exp concurrent validity or exp construct validity or exp content validity or exp predictive validity or exp test validity} and{course work assessment or exp academic achievements or exp educational assessment or exp profiles or exp portfolio or exp classroom observation} and {exp tests or exp certification or baseline tests or foundation tests or selection tests of graduations tests} and {exp schools or pre-school or exp British infant schools, etc.}.

Reliability/validity	Assessment by teachers	Summative purpose	Relevance to school
Reliability (all forms) Validity (all forms) Dependability Consistency Comparability Moderation Quality assurance	Teacher assessment (asst) Teacher-based asst Course work asst Ongoing asst School-based asst Classroom asst Embedded asst Profile Portfolio Observation Process asst	Summative assessment Examination Certification National tests (ing) Tests (ing) Baseline tests/assessment Foundation tests/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

APPENDIX 2.3: Journals handsearched

Journal	Type of search	Searcher	Dates searched	No of articles
American Educational Research Journal	Hand	WH	1988- 2002	3
American Journal of Evaluation	Online	WH	1998 - 2003	0
Assessment in Education	Hand	WH	1994 - 2003	13
British Educational Research Journal	Hand	WH	1987 - 2003	3
British Journal of Educational Psychology	Online	WH	1999 - 2003	4
British Journal of Educational Studies	Online	WH	1999 - 2003	0
British Journal of Educational Technology	Online	WH	1999 - 2003	0
Cambridge Journal of Education	Hand	WH	1988 - 2003	1
Curriculum Journal	Hand	WH	1988 - 2003	0
Educational Assessment	Online + Hand	WH	1990 - 2002	8
Educational Evaluation and Policy analysis	Hand	WH	1988 - 2002	0
Educational Measurement	Hand	WH	1993 – 2002	5
Educational & Psychological Measurement	Hand	WH	1995 - 2002	9
Educational Research	Hand	WH	1980 - 2002	3
Educational Researcher	Hand	WH	1985 - 2003	2
Educational Review	Hand	WH	1991 - 2002	0
Educational Studies	Online + Hand	WH	1993 - 2000	2
Educational Studies in Mathematics	Online	WH	1996 - 2003	0
Journal of Curriculum Studies	Hand	WH	1990 - 2003	0
Journal of Educational Measurement	Hand	WH	1990 - 2002	0
Journal of Education Policy	Hand	WH	1987 - 1996	1
Journal of Educational Psychology	Hand	WH	1985 - 2003	0
Oxford Review of Education	Hand	WH	1985 - 2002	0
Research Papers in Education	Hand	WH	1986 - 2003	0
Studies in Educational Evaluation	Hand	WH	1986 - 2003	5
Teachers College Record	Hand	WH	1995 - 2002	0

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 Reliability/validity reported
- B.1.1 Reliability
- B.1.2 Validity

B.2 Comparison of teacher assessed outcomes with other form of test/exam/assessment
- B.2.1 Yes
- B.2.2 No

or with external rating/scoring
- B.2.3 Yes
- B.2.4 No

B.3 Achievement assessed
- B.3.1 Reading
- B.3.2 Writing
- B.3.3 English
- B.3.4 EFL/EAL
- B.3.5 Maths
- B.3.6 Science
- B.3.7 Practical maths/science/tech
- B.3.8 Art/Music/Dance/PE
- B.3.9 Extended project
- B.3.10 Other (Please specify.)

B.4 Students' teacher assessed tasks
- B.4.1 Set externally
- B.4.2 Set by teacher to external criteria
- B.4.3 Portfolio
- B.4.4 Regular work/embedded assessment
- B.4.5 Project

B.5 Type of scoring
- B.5.1 Teacher marking/grading using external criteria
- B.5.2 Teacher marking/grading using own criteria
- B.5.3 Students marking/grading, moderated by teacher
- B.5.4 Part external

(Note: If scoring is wholly external, exclude in accord with exclusion criterion C.)

B.6 Assessment purpose
- B.6.1 Baseline (foundation)
- B.6.2 Certification
- B.6.3 National/Regional/State assessment
- B.6.4 Informing other teachers (transfer)
- B.6.5 Informing parents/students
- B.6.6 Selection
- B.6.7 Formative
- B.6.8 Accountability
- B.6.9 Monitoring
- B.6.10 Other (e.g. research)

APPENDIX 3.1: Systematic map of keyworded studies

Country in which the studies were carried out: Table A3.3.1 gives the countries in which the studies were carried.

Table A3.1.1: Country

Country	Number of studies
Australia	1
England	15
Greece	1
Israel	1
USA	12
Total	***30***

Topic focus: All studies were categorised as focusing on assessment. Although they generally concerned assessment of some curriculum subject, this was not indicated under the topic focus but recorded as one of the review-specific keywords (see Table A3.1.7).

Population focus: Since all studies were concerned with the assessment of learners, learners were indicated as a population focus even though the study might have been primarily concerned with teachers' judgements rather than pupils' performance. Twenty-five studies were classified as having teaching staff as the focus.

Age of learners: Table A3.1.2 give the age of learners. It shows the combination of age ranges included in the studies. For instance, of the 17 studies concerned with students aged 5 – 10, three also included students aged below four and six also included students in the age range 11 – 16.

Table A3.1.2: Age of learners (not mutually exclusive)

Age of learners (years)	0-4	5-10	11-16	17-20	Total
0-4	3	3	1	0	3
5-10	3	17	6	0	17
11-16	1	6	14	1	14
17-20	0	0	1	6	6

Gender of learners: All studies involved learners of both sexes.

Educational setting: Table A3.1.3 shows that the majority of studies were set in state primary and secondary schools. Some studies involved learners from more than one educational setting.

Table A3.1.3: Educational setting (categories not mutually exclusive)

Educational setting	Number
Independent school	2

Educational setting	Number
Nursery school	3
Post-compulsory education institution	1
Primary school	19
Secondary school	13
Other educational setting (sixth form college)	1

Type of study: The majority of studies were designed to compare assessment by teachers with some other assessment of the learners, thus the majority were 'exploration of relationships'.

Table A3.1.4: Type of study (mutually exclusive categories)

Type of study	Number
Description	1
Exploration of relationships	18
Evaluation: Naturally-occurring	8
Evaluation: Researcher-manipulated	3

Review-specific keywords

Focus on reliability and/or validity of teacher assessed measures 12 studies focused mainly on reliability, 18 mainly on validity, with two reporting data for both, but were mainly concerned with reliability.

Table A3.1.5: Reliability/validity focus (mutually exclusive categories)

Main focus	Number
Reliability	12
Validity	18

Comparison of teacher assessed outcome with a test/examination: 21 studies involved such comparison and nine did not.

Comparison of teacher assessed outcome with an external rating or scoring of the same evidence: Six studies involved such comparison and 24 did not. Table A3.1.6 shows the number of studies in which comparisons of different kinds were made.

Table A3.1.6: Studies in which comparisons were made between teacher assessments and another test or examination and/or with external ratings of the teacher assessed evidence

Comparison with external rating/scoring	Comparison with other form of test/examination		Totals
	Yes	No	
Yes	3	3	6
No	18	6	24

Aspect of achievement/subject assessed

Table A3.1.7: Aspects of achievement assessed (categories not mutually exclusive)

Achievement assessed	Number
Reading	13
Writing	8
English	8
Maths	20
Science	15
Practical maths/science/tech	5
Art/music/dance/PE	1
Other (foreign language, verbal reasoning, history, geography, Latin, bible studies)	10

The origin of the teacher-assessed task: This refers only to the tasks that were assessed by teachers, not to tasks or tests, if any, used to compare with the teacher assessment.

Table A3.1.8: Origin of the teacher-assessed task (categories not mutually exclusive)

Students' teacher assessed tasks	Number
Set externally	4
Set by teacher to external criteria	5
Portfolio	3 (inc 1 also regular work)
Regular work/embedded assessment	18 (inc 1 also portfolio, 1 set by teacher to external criteria)
Project	2

Type of scoring of teacher-assessed tasks: Again this refers only to the tasks that were assessed by teachers, not to tasks or tests, if any, used to compare with the teacher assessment

Table A3.1.9: Type of scoring of teacher assessed tasks

Type of scoring	Number
Teacher marking/grading using external criteria	27
Teacher marking/grading using own criteria	3
Student marking, moderated by teacher	0
Part external	0

Purpose(s) of the teacher assessment

Table A3.1.10 gives the number of studies concerned with assessment of different purposes. Several related to more than one purpose, as indicated in Table A3.1.11.

Table A3.1.10: Purpose(s) of the teacher assessment (categories not mutually exclusive)

Purposes of assessment	Number
Certification	6
National/Regional/State assessment	12
Informing other teachers (transfer)	1
Informing parents/students	6
Selection	1
Formative	3
Accountability	2
Monitoring	4
Other (e.g. research)	7

Table A3.1.11 shows the number of studies that served more than one purpose. For example, of the 12 studies which served national or State assessment requirements, one was also used to inform parents, one served a formative purposes and two were categorised as concerned with monitoring.

Table A3.1.11: Combinations of purposes of the teacher assessment (categories not mutually exclusive)

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	Other
(a) Certification	6	0	0	0	0	0	0	0	0
(b) National or State assessment	0	12	0	1	0	1	0	2	0
(c) Informing other teachers (transfer)	0	0	1	1	0	0	0	1	0
(d) Informing parents/students	0	1	1	6	0	1	1	2	2
(e) Selection	0	0	0	0	1	0	0	0	0
(f) Formative	0	1	0	1	0	3	1	0	1
(g) Accountability	0	0	0	1	0	1	2	0	0
(h) Monitoring	0	2	1	2	0	0	0	4	0
Other (e.g. research)	0	0	0	2	0	1	0	0	7

Relationship between categories

Table A3.1.12 shows the number of studies in which there were various combinations of teacher assessed tasks and type of scoring. Tables A3.1.13 and A3.1.14 show how aspects assessed and types of study varied with the educational setting of the studies. Tables A3.1.15 to A3.1.17 show how the areas of achievement assessment, the type of tasks and the type of scoring varied with the purpose of the assessment.

Table A3.1.12: Number of tasks of different kinds scored in various ways

Teacher assessed tasks (categories not mutually exclusive)	**Type of scoring** (mutually exclusive categories)		
	Teacher marking/grading using external criteria	Teacher marking/grading using own criteria	Students marking/grading, moderated by teacher
Set externally	4	0	0
Set by teacher to external criteria	5	0	0
Portfolio	3	0	0
Regular work/embedded assessment	15	3	0
Project	2	0	0

Table A3.1.13: Aspects assessed in different educational settings (categories not mutually exclusive)

Type of educational setting (categories not mutually exclusive)	**Reading**	**Writing**	**English**	**Maths**	**Science**	**Practical maths/ science/ tech**	**Art/music /dance/PE**	**Other**
Independent school	1	1	1	2	1	0	0	1
Nursery school	3	2	1	3	0	0	0	0
Primary school	13	8	5	16	8	2	1	3
Secondary school	2	1	3	6	7	3	0	7

Table A3.1.14: Types of study in different educational settings

Type of educational setting (categories not mutually exclusive)	**Description**	**Exploration of relationships**	**Evaluation: naturally-occurring**	**Evaluation: researcher-manipulated**
Independent school	0	1	1	0
Nursery school	0	2	1	0
Primary school	1	9	7	2
Secondary school	0	9	3	1

Table A3.1.15: Aspects of achievement assessed for different purposes (categories not mutually exclusive)

Achievement assessed	**Assessment purpose**									
	Baseline (foundation)	Certification	National/ Regional/ State assessment	Informing other teachers (transfer)	Informing parents/ students	Selection	Formative	Accountability	Monitoring	Other (e.g. research)
Reading	0	0	6	1	4	0	0	0	4	5
Writing	0	0	4	0	3	0	1	0	3	3
English	0	2	5	0	0	0	1	0	1	1
EFL/EAL	0	0	0	0	0	0	0	0	0	0
Maths	0	2	10	1	4	0	1	0	4	6
Science	0	4	8	0	0	1	2	1	1	1
Practical maths/science/tech	0	2	2	0	0	0	1	1	1	0
Art/music/dance/PE	0	0	0	0	1	0	1	0	0	1
Extended project	0	0	0	0	0	0	0	0	0	0
Other (foreign language, verbal reasoning, history, geography, Latin, bible studies)	0	4	0	1	2	1	0	1	1	3

Table A3.1.16: Type of teacher-assessed tasks for different purposes (categories not mutually exclusive)

	Baseline (foundation)	Certification	National/ Regional/ State assessment	Informing other teachers (transfer)	Informing parents/ students	Selection	Formative	Accountability	Monitoring	Other (e.g. research)
Set externally	0	1	3	0	0	0	0	0	0	0
Set by teacher to external criteria	0	1	2	0	2	0	1	1	1	1
Portfolio	0	0	3	0	1	0	1	0	1	0
Regular work/embedded assessment	0	3	6	1	3	1	1	0	2	6
Project	0	1	0	0	0	0	1	1	0	0

Table A3.1.17: Type of scoring of teacher-assessed tasks for different assessment purposes

	Baseline (foundation)	Certification	National/ Regional/ State assessment	Informing other teachers (transfer)	Informing parents/ students	Selection	Formative	Accountability	Monitoring	Other (e.g. research)
Teacher marking/ grading using external criteria	0	5	12	0	5	0	3	2	3	7
Teacher marking/ grading using own criteria	0	1	0	1	1	1	0	0	1	0
Students marking/grading, moderated by teacher	0	0	0	0	0	0	0	0	0	0
Part external	0	0	0	0	0	0	0	0	0	0

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

Study	Focus	Achievement assessed	Study type	Evidence weight
Abbott *et al.* (1994) Some sink, some float: National Curriculum assessment and accountability	Reliability	Science Practical maths/science/ tech	Evaluation: naturally-occurring	Medium

Aims
To investigate the question: Are standard tasks (the authors refer to them as SATs) more reliable than teacher assessments (TA) made during the ordinary course of events?'

Study design
Descriptive observations of SAT administration in year 2 classes in three schools, followed by interviews with teachers after the administration. The observed events were not intentionally influenced by the observation.

Data collection
Data collection took place in three classrooms in three different schools, during the administration of the standard task in science (Floating and Sinking) for seven year-olds in 1991. 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
A straightforward account of the observations of the administration of SATs in the three classes, in narrative and tabulated form.

Authors' findings
There were considerable variations among the schools in relation to the conditions - interruptions, help, etc. While variations in conditions were obvious to anyone, 'continuous observation while the SAT took place revealed factors in teachers' presentations of the tasks which made the Government's declared aim of standardising their assessment techniques and children's experience appear completely out of reach' (p 163). The instructions, which set out to be precise, detailed, and leaving little room for individual interpretation, were mediated through different teachers' priorities, concerns and pedagogic styles. While none of the teacher set out to distort the results, 'their practice diverged greatly in several dimensions: their interpretation of 'leading' children, their readiness to accept pupils' answers as evidence of ability to interpret findings by linking one factor with another or by making a generalisation' (SEAC, 1991, p 67) and the time they allowed.

The number of children assigned to each level varied across the schools. 'Evidence from observation suggested that children at Leigh, where no Level 3 grades were recorded, probably achieved as much as those at Greenside' (where six out of 10 children were given level 3 (p 166).

Author's conclusions
Although the results showed that the standard tasks were 'extremely unreliable' (p166) the

authors questioned the decision of the Government to drop the SAT from the testing programme for the next and subsequent years. They pointed out that the skills assessed during the standard tasks in science in 1991 were valuable ones which were not addressed in any of the 1992 science tasks which focused on describing, grouping and comparing objects and materials and on ways of discussing and recording weather conditions.

The authors argue that the standard task activities are 'perfectly valid and useful lessons, but hardly useful means of assessment' (p 168). They suggest that the Floating and Sinking task could have been abbreviated and that the worksheets which took children so long to complete provided no useful evidence of different levels of achievement.

The authors also point out that teachers found other Standard Tasks equally difficult to use (e.g. the reading tasks). 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' (p 171).

Study	Focus	Achievement assessed	Study type	Evidence weight
Bennett *et al.* (1993) Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill	Validity	Reading English Maths	Exploration of relationships	High

Aims
To test a model relating to tested achievement, gender, teachers' behaviour perceptions and teachers' judgements of academic skill.

Study design
Entire populations from four schools were involved. Three schools were in Cleveland, Ohio, and one in the Bronx, New York. 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 given for effort and conduct were given by the teacher and 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 maths, handwriting and reading comprehension, following common criteria. Means of these were used.

Grades were given in June for report cards, across spelling, phonics, reading, maths, handwriting and English (not all for the Kg). Average grades were computed.

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'.

Regression analyses were run separately for each grade within each district, thus permitting both location and grade to be treated as replications.

Authors' findings
Gender differences: Girls received consistently significantly higher behaviour grades than boys for grades 1 and 2. There were no gender differences in academic test scores and academic grades. For academic ratings, there was a gender difference (in favour of girls) only in Grade 1.

From the path coefficients (standardised partial regression weights): The Kg behaviour grade consistently affected teachers' academic judgements after controlling for gender, academic score and missing data. Effect sizes were large.

Kg test scores consistently and significantly affected academic grades (less so academic 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. Further, behaviour grade had a consistent direct effect on academic judgement (after controlling 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’s conclusions
‘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 1.0 SD change in tested academic skill produced a shift in academic appraisal for grades 1 and 2 only marginally larger than that for behaviour; 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.

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’ (p 353). Second, there is‘the need for more concerted effort toward making teachers aware of the potential influence of student behaviour on their academic appraisals’ (p 353).

Study	Focus	Achievement assessed	Study type	Evidence weight
Brown *et al.* (1996) The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations	Validity	Practical maths/science/ tech	Exploration of relationships	Medium

Aims
To examine the relationship between practical skills and those tested in the theory papers in three 'A' level science subjects. This contributes to the construct validation of practical assessment in these subjects.
To identify the extent to which the practical skills specified in these science subjects are dependent on the context in which they are set.

Study design
Secondary data analysis of assessment and exam results using multiple correlations and factor analysis. Samples were selected of candidates for biology, physics and chemistry 'A' level. Data were collected about their performance on theory papers and on teacher-assessed practical skills. For investigating construct validity, intercorrelations among theory and practical components were computed separately for each science group and the same data also subject to factor analysis.

Data collection
Data were generated by teachers and the examination board and collected from the examination board by researchers. The teacher assessment of practical skills was carried out by teachers for their own students during their 'A' level courses on two occasions. The skills specified were three for biology (A – C), four for chemistry (A – D) and five for physics (A – E).

Data analysis
Intercorrelations (Pearson correlation coefficient) and factor analysis loading (varimax) were obtained from the data for each subject and for each set of candidates taking pairs of subjects (since there were insufficient taking all three sciences). Evidence for convergent and discriminant validity was based on comparative magnitude of the correlations and on the factor structure.

Authors' findings
(1) Intercorrelations between practical skills are lower than between theory exams (Table II p 383).
(2) In biology and physics, mean intercorrelations between theory and practical skills are lower than the intercorrelations of the individual tests. In chemistry, the intercorrelation between the theory and practical skills is higher than the intercorrelation of the practical skill results alone. The authors say that this suggests lower construct validity of the practical skill assessment in chemistry (Table II, p 383).
(3) In the factor analysis, practical skills A and B consistently load higher onto a different factor from the theory assessment scores. Practical skills C and D tend to load equally onto the same factor scores as the theory tests and the practical tests. The authors argue that this suggests that there is some evidence of a practical construct being tested by each of the skills A and B but less so for C and D (tables III, IV, V, VI and VII, pp 384-386).
(4) There is a low correlation between skills tests A & B in the different subjects but slightly higher correlation for skills test C. The authors suggest that this indicates some evidence for context independence for skill C (tables VIII and X, pp 386-387).
(5) On the factor analysis of the skills scores (tables IX and XI), the skills tests in each of the

subjects load onto two different factors. The authors suggest that the weight of evidence is that both groups of practical skills are strongly weighted to contexts (subject)(p 388).

Authors' conclusions
The authors conclude that 'some evidence of construct validity has been shown, but to varying degrees in each subject. Hence teacher assessment of students' practical work seems to make a valid contribution to describing and quantifying attainment in these subjects. There is less evidence, however, that different facets of practical work make extensive contributions. There may be several reasons for this. Teacher assessment may be subject to a halo effect in which an over-arching impression of a candidate's quality leads to similar judgements being made of performance in the different skills. Or it is possible that the skills which have been identified for each subject are intrinsically interrelated and scores on them will inevitably be highly correlated' (p 388).

There was little evidence of the generalisability of skills assessment across subjects, indicating that the skills assessment was context dependent. They conclude that this suggests that what was being assessed was subject rather than skill domains and that the results indicate a need for continued assessment of skills within subject domains.

Study	Focus	Achievement assessed	Study type	Evidence weight
Brown *et al.* (1998) An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England	Validity	Science Practical maths/science/ tech	Exploration of relationships	High

Aims
To compare two different methods of assessing the project which is part of the 'A' level examination in biology
To identify any effects of changes in the methods of assessment (from external rating to rating by teachers) on the construct validity of the project and on its contribution to the overall examination

Study design
The results of two samples of A-level students were compared in terms of intercorrelation between theory and project papers and the outcomes of factor analysis of their results.
Two data sets were drawn for the analysis, one from the 1993 examination, when the project was externally assessed by the Examination Board, the other from that of 1996, when the project was assessed by teachers. Both samples were roughly representative of the subject entry in type of school, size of entry and geographical location. They constituted about 10% of the total entry. The input data consisted of candidates' scores on each of the theory components, on 13 process sections of the project for the 1993 data and, for the 1996 data, on four teacher-assessed (TA) scores on the project (planning, implementing, interpreting and concluding and researching). In both years, teachers were also required to make an assessment of candidates' practical skills.

Data collection
Data were extracted from the examination results for the marks given in the three theory papers, the marks given for the various aspects of the project and the practical skills as assessed by the teachers from observations during the course of laboratory work.

Data analysis
Construct validity was determined by correlational and factor analysis (principal component with rotation to varimax criterion) of candidates' scores using the SPSS package. Factor analysis was conducted after the data had been scrutinised for appropriateness, using the Kaiser-Mayer-Olkin measure of sampling adequacy and Bartlett's test of sphericity.

Authors' findings
Mean intercorrelation between the three theory papers for 1993 was 0.78, while intercorrelations between the theory papers and the project was 0.45, suggesting that something quite different was assessed by the project in comparison with that done by the theory papers. Teacher-assessed practical skills (which teachers were asked to give in all years) formed a separate factor.

Intercorrelation between the project section scores and all other components were all positive and varied widely. For 1993, factor analysis showed a clear theory factor with very low loading of the project components. For the 1996 examination, the mean intercorrelation among the theory papers was 0.84 and between theory and project 0.55. Factor analysis found two factors. Factor 1 was a theory factor but also two of the teacher-assessed project skills loaded significantly onto it. Also, the teacher-assessed practical skills loading were different from the

1993 findings 'This suggests that the TA procedure no longer assessed a construct different from theory and the project'.

Authors' conclusions
The authors point out that the 1996 examination was the first in which teachers were asked to assess the project, but, as they were used to assessing pupils' work (and were already doing this for the practical skills), there was unlikely to be a novelty factor influencing the results. 'The 1993 data showed that, overall, the project demonstrated construct validity in that it tested something that was different from the objectives tested by the theory papers' (p 94).

'Of considerable interest was that the project factors received very low loading from the scores on practical skills derived from continuous assessment over the duration of the course.'

The conclusion is that the two forms of practical - the ongoing skills during the course and the project - tested different constructs from each other and from the theory papers.

For the 1996 examination, 'we find that the evidence for construct validity is much less compelling. Seen from the perspective of the teachers who prepare candidates for this examination, they were required to assess the candidates' practical skills given a set of criterion-based descriptors. Four skills were assessed in the project, two of the same skills on a minimum of two occasions as part of the ongoing practical/lab work. The outcome of these requirements was that the assessment in 1996 became more similar to the theory assessment than they had been in 1993 ... A speculative suggestion to explain this might be that a halo effect operated. Having assessed the practical abilities of their candidates over two years ...why would they expect different performances to emerge from the same candidates on the project?' (p 94).

Study	Focus	Achievement assessed	Study type	Evidence weight
Chen and Ehrenberg (1993) Test scores, homework, aspirations and teachers' grades	Validity	Reading Maths Bible	Exploration of relationships	Low

Aims
To clarify the direct effects of aspiration, homework and scores of standardised tests on teachers' grades and the effects of exogenous variable, SES and gender on the endogenous variables, aspirations, homework and achievement test scores.

Study design
The study tested a path model connecting eight variables by subjecting the data to the CALIS procedure (in the SAS package) to explore linear relationships. It was hypothesised that the variance of the dependent variable (teachers' grades) is explained only by the endogenous variables and that variance of the endogenous variable is explained by the exogenous variable (an over-identified model). Path analysis was used to test the research hypothesis.

Data collection
The study was conducted on a heterogeneous sample of sixth-grade students in a medium-sized, affluent town in Israel.

Information about their background was collected from students by personal questionnaire. Information about achievement was obtained from standardised multiple-choice tests based on the formal school curriculum that sixth graders are supposed to know at the end of the elementary school.

Students' aspirations were measured by the average scores in answer to the following:
(a) English: I want to study in (1) the first ability group, (2) the second ability group or (3) another ability group.
(b) Mathematics. as for English
(c) Do you plan to attain a matriculation certificate: range from 1 (certainly) to 4 (certainly not). Regular preparation of homework was rated by the teachers on a seven-point scale, ranging from 1 (doesn't prepare homework) to 7 (always prepares homework).

Teachers' grades were on the average grades on a student's report-card in reading, bible and maths at the end of sixth grade, varying from 4 (weak) to 10 (excellent).

Data analysis
The model specified by the researchers was tested 'by using the linear structural relationship model by means of the CALIS (equivalent to the LISREL) procedure in the SAS package. This provides intercorrelations among the variables. The fit of the model is judged by the size of the residual difference between the observed correlations and the reconstructed correlation matrix'.

Authors' findings
'The strongest direct influence on teachers' grades was homework.' (p 413) Achievement scores was the next strong influence on teachers' grades. There was a very small, but statistically significant influence of aspirations - those students with higher aspirations receive slightly higher evaluations as a result of these aspirations alone (all other variables being controlled).
The strong influence of homework and achievements on tests are in accordance with the hypotheses. 'It is, however, surprising that the influence of homework is much greater than the influence of test achievements (b = 0.594, b = 0.346). There was no indication of influence of

background (exogenous) variables on teachers' grades.' (p 413) (There are additional findings linking aspirations, homework and test achievements.)

Authors' conclusions
The study confirms the hypothesis regarding the over-identified path model, which explains 72% of the variance in the teachers' grades. It suggests that teachers grade their students properly according to their knowledge of the subject matter, their efforts and aspirations only. There is no indications that they take into consideration irrelevant factors, such as gender or SES. However, the excessive importance attributed to homework indicates indirect preference for students of high SES and female students (since homework is influenced by SES, aspirations and gender).

'This relationship can be explained by the realities of the elementary school in Israel. Homework is an individual method of learning in which students advance in their studies according to their own pace of learning. On the other hand achievements reflect a standardised mastery of subject matter, regardless of the teacher instruction... Achievement tests don't provide the proper information about students' behaviour, study habits, diligence and organization and different cognitive approaches. It is probable that teachers assume homework to be a more valid indicator of the student's knowledge and study habits than on-tie standardised tests' (p 414).

'Another possible explanation ...is related to the teachers' interest in 'industrial peace' in the classroom. By rewarding conforming students, who are disciplined and regularly prepare their homework, they achieve a positive school climate.. This may sometimes require ignoring the real achievements of non-cooperative students.... it ensures a convenient school climate at the expense of the very able students who make the grade with a little homework' (p 415).

The overweighting of homework compared with test achievements requires some consideration. It is possible that such a state of affairs is detrimental to the able but non-conforming, as well as to students of low SES.

Study	Focus	Achievement assessed	Study type	Evidence weight
Coladarci T (1986) Accuracy of teacher judgements of student responses to standardised test items	Validity	Reading Maths	Exploration of relationships	High

Aims
To study whether teachers can correctly gauge student responses item by item on a valid achievements test that had been administered concurrently to their students.

Study design
Teachers gave their estimates for the performance of six randomly selected students on each of the test items. The students had taken the actual tests two weeks earlier, but the results were not known to the teachers.

Data collection
Teachers of third- and fifth-grade students gave their judgements of the students' achievement in interviews in which the interviewer randomly selected six students (two from each ability group). For each student, the interviewer asked the teacher to indicate whether he or she thought the students correctly answered specific items on the SRA test. This was done for each item on the Reading Vocabulary, Reading Comprehension, Mathematics Concepts and Mathematics Computation sub-tests for third and fifth grade (Form 1).

SRA Achievement Series 1978 tests were used.

Data analysis
Item-level results were summed to form sub-test results, total reading, total mathematics and total test results. This was the same for the SRA tests score and the teacher's judgements. Descriptive statistics were calculated for the student-performance and teacher-judgement measures.
Correlations of the aggregate measures of teacher judgements and aggregate measures of teacher judgement were given.

Correlations were computed between each performance/judgement agreement measure and three indicators of achievement: (a) the same measure on which performance/judgement agreement was established, (b) the student's total score across the four sub-tests (i.e. total test), and (c) the student's general designation provided by the teacher at the outset of the study (‘below’, ‘at’, or ‘above’ grade level).

Performance/judgement agreement measures were computed from the percentages of correct judgements for each item averaged over the tests. Intercorrelations of these agreement measures were also computed.

Author’s findings

(a) Aggregate measures of teachers' judgements of their students' responses to items on a standardised achievement test correlated positively and substantially with aggregate measures of students' actual responses (range 0.67 to 0.85).
(b) Teachers accurately judged their students' responses to individual items.
(c) The accuracy of teachers' judgements varied significantly as a function of sub-test.
(d) There were significant individual differences among teachers in the accuracy of their judgements.

(e) Teachers were least accurate in judging low-performing students and most accurate in judging high-performing students (p 144).

For all sub-tests, some students were judged correctly for fewer than half of the test items, whereas other students were judged correctly for nearly all the items. The reason for this was a combination of teacher effect and student effect. A one-way ANOVA with teacher as the grouping factor found a significant teacher effect for ‘maths concepts' but not, it is assumed, for other sub-tests.

Author’s conclusions
The findings in relation to greater accuracy of teacher judgements for high-achieving students is expected. (Teachers would be more likely to report that the student would get the item correct and students more likely to succeed and there would be relatively few events to the contrary.) No simple response set would work for students further down the achievement scale. 'For the moderate and low-achieving student, teachers doubtless realised that there were many items that the student would not answer correctly. What was difficult was to decide where the errors would occur. And the lower the student's proficiency, the more difficult - and inaccurate - this judgement was. These results point tentatively to the disturbing implication that students who perhaps are in the greatest need of accurate appraisal made by the teacher in the interactive context are precisely those students whose cognition has a greater chance of being misjudged' (p 145).

The observed relationship between teacher accuracy and task can be explained, in part, by (a) the degree to which teachers provide direct instruction in the task domain and (b) the amount of information teachers have that bears on student proficiency in that domain. In mathematics, typically, there is more direct instruction in computation than in concepts. Another factor might be the complexity of the task. (The author suggests that understanding would be clarified by using test items written in an open-ended format rather than multi-choice.)

In conclusion, ‘applied to the interactive decision making, these results suggest that the accuracy of a teacher's judgement is influenced by characteristics of the teacher, student and academic task' (p146).

Study	Focus	Achievement assessed	Study type	Evidence weight
Crawford *et al.* (2001) Using oral reading rate to predict student performance on state-wide achievement tests	Validity	Reading Maths	Exploration of relationships	Medium

Aims
The aim of the study is to expand on previous research by exploring the ability of various reading rates to predict eventual performance on state-wide achievement tests. Establishing a critical range of reading rates will extend the use of curriculum based measurement (CBM) as a viable classroom tool for monitoring students' progress toward meeting state-wide benchmarks in mathematics. This purpose is supplemented with an interest in extending the predictive power beyond a single year (p 307).

Study design
One group of students was followed for two years, during which CBM reading results were collected and, at the end, state-wide test results collected.

Data collection
To assess reading rates, three passages were chosen for use during each year of the study. Each were modified to have approximately 200-250 words and to have cogent beginnings and ends. Passages were taken from the Houghton Mifflin Basal Reading series which was used in the school district and participating school. Each passage was orally read and timed for one minute; this was then scored for correct words and errors.

Third-year students were tested on state-wide maths and reading assessments; both were criterion-referenced tests, containing multiple-choice questions and performance tasks. Results were reported on standardised scores in a Rasch scale.

Data analysis
Three types of outcomes are reported. 'First, descriptive statistics are reported for Year 1 and Year 2. Second, correlations between oral timed readings and the state-wide reading and maths tests are reported. Third, chi-square analyses are presented, allowing us to determine which levels of oral reading rates are most predictive of performance on the state-wide tests' (p 314).

Author's findings
(a) Results show that the mean scores on the state-wide reading assessment met the state-established criterion for a passing score. However, the mean scores on the state-wide maths assessment fell short of the established criterion by two points. 65% passed the reading assessment and 45% passed the maths assessment.
(b) There was a large increase in the number of correct words read per minute between second and third grade.
(c) The mean gain in oral reading rate was approximately 42 correct words per minute. There was no relationship between the initial reading rate and amount of gain.
(d) There was a strong relationship between second- and third-grade oral reading rates.
(e) There was a moderate correlation between second-grade timed oral readings and state scores (on reading test), slightly higher than those obtained in the third grade.
(f) There was a moderate correlation between performance on the math test and timed oral readings, with the across years correlation slightly higher than the within years correlation.
(g) Of 37 students reading in the top three quartiles, 29 passed the state reading test, whereas only 29% of the students reading in the first quartile passed. Of the students reading at least 72 correct words per minute in second grade, 100% passed the state-wide reading test in

third grade. A chi-square representing second grade reading rates and state-wide test scores, demonstrated statistical significance.

(h) Within-year data failed to generate definitive patterns between rates for reading and scores on the state-wide maths test.

(i) No statistical significance was seen for between-years data on rates for reading and state-wide maths achievement.

A Pearson correlation coefficient was calculated between second and third grade reading rates, revealing a strong relationship (r = 0.84) Reading rates increased substantially between years. There was no correlation between improvement and second grade reading rate.

Correlations between the state-wide reading assessment in the third year and reading rates in the second and third year were moderate and slightly higher for the second year than the third year. Correlations between the state-wide maths assessment and the reading rates were smaller and again larger for the second grade than the third grade. There were no significant differences between the within -year and across-years correlations.

Using chi-square, the reading rates/test performance correlations were shown to be significant.

Authors' conclusions

The longitudinal data presented in this study demonstrate that CBMs are sensitive enough to detect growth for almost every student, with 50 out of 51 students in this study improving their rate of reading over the course of one year. CBM procedures also seemed to lack bias in that the gains students made on the measures were not an artefact of their starting points, as no significant differences were found between the amount of gain made by students who had low initial rates and those that had high initial rates. The authors claim that the results demonstrate that teachers can rely on the accuracy of CBMs in monitoring the reading progress of all students.

There are obvious benefits for teachers who use CBM in reading to monitor students' progress such as the ability to predict students' future performance on state-wide tests. Perhaps the most important finding of this study is the fact that 100% of the second-grade students who read at least 72 correct words per minute passed the state-wide reading test taken the following year. These clear and simple data communicate powerful information to teachers (p 320).

Study	Focus	Achievement assessed	Study type	Evidence weight
Delap MR (1994) An investigation into the accuracy of A-level predicted grades	Validity	English Maths Science History Geography Sociology	Exploration of relationships	Medium

Aims
To determine the accuracy of predicted grades supplied to UCCA by way of the referees' reports and to analyse these data in the light of previous studies (p 136).

Many of the previous studies concentrated upon determining the accuracy of predicted grades once the data have been aggregated within particular categories (e.g. subject and examining board). This method has been followed for the initial part of the data analysis, using the categories of gender, centre type, ethnic origin, examining board, age and subject. The second part of this report presents analysis of the data which attempts to evaluate the influence of each category once the influence of the remaining categories are taken into account (p 136).

Study design
Relationships between predicted and actual grades were explored using aggregated data for each variable, with no attempt to take account of the interaction between the categories. In a second analysis a multi-level approach enabled the interrelationships among categories to be taken into account.

Data collection
UCCA application forms and 'A' level examinations

Data analysis
1. Summary statistics to describe the data
2. Pearson product moment correlation coefficient
3. Regression analysis, based on a two-level hierarchy with applicants at level 1 nested within subjects at level 2

Author's findings
Descriptive results (cautious)
(a) There was a correlation between the predicted and actual results (Pearson's 0.66). Predicted grades tended to over-estimate by, on average, one point or half a grade.
(b) Gender: The male applicants appear to have performed slightly better than female applicants, but there was little difference between the percentage of pessimistic, accurate and optimistic grades.
(c) Centre: Further education colleges showed the largest difference between the means of the predicted and actual grades, with the lowest proportion of accurate predictions and highest proportion of very optimistic predictions. This was in marked contrast to those of independent and selective schools.
(d) Ethnicity: A greater proportion of predictions for White ethnic origin were classified as accurate than for applicants of Asian ethnic origin.
(e) Age: Predicted grades became steadily less accurate as the age of the applicant increased, while over a half were accurate for 1seven year-olds, only a quarter were accurate for applicants aged over 20.

(f) Examining boards: Predictions for examining board W were the most pessimistic, while those of board X were the most accurate. 20% of applications for the other boards were very optimistic.
(g) Subject: There was wide variation between subjects in the proportion of accurate predicted grades ranging from 38.6 to 51.3. The mean difference in actual and predicted grades was smallest in French but largest in history.

Regression responses (more sure)
(h) Almost all the variance between predicted and actual grades was attributable to the applicant level variables with very weak effects at the subject level. When the actual grades were included as explanatory variables, the variance decreased.
(i) Despite the appearance of the data on gender, there is a slight gender effect, with predictions for females being slightly less optimistic than those for males
(j) The predictions for further education establishments are significantly different from those for secondary comprehensive schools.
(k) The mean difference between predicted and actual grades is not influenced by ethnic origin of the applicant.
(l) There is a significant difference between the optimism of predictions for applicants aged 19 and those aged 18, while the difference for candidates who are 20 is almost significant. Predictions for applicants aged 19 are slightly less optimistic that those of applicants who are 18
(m) Predictions for French, chemistry and geography are significantly less optimistic than those for physics.

Author's conclusions
In many respects, the aggregated results of the study are similar to the findings of previous studies. The mean difference between predicted and actual grades is approximately half a grade and there is apparently a significant correlation between predicted and actual grades. However, the results of this study have highlighted some of the shortcomings in previous analyses and the differences that can occur if other variables are not taken into account in the analyses.
The results of the two-level model reveal that predictions for UCCA applicants aged 19 are on average slightly more accurate than those for 17, 18 year-olds since they are slightly less optimistic. There is also a small, but significant, gender effect: predictions for males are more optimistic (less accurate) than those for females. Some evidence was also found to support the view that the predicted grades from further education establishments are more optimistic on average then those from other types of centre. Finally, the origin of the applicant has been shown to have no influence upon the optimism or pessimism of teachers predicted grades.
'It has been demonstrated that unless the distribution of the actual grades for each of the sub-categories being compared are similar, the interpretation of the data will be misleading. For example one may have been led to conclude from the summary analysis that there significant differences between the accuracy of prediction for some of the examining boards.' (p 147) The finding from the multilevel analysis reveals that while the difference exists in the raw data, once other factors are taken into account – i.e. gender, final grade, centre type and subject - there is no significant difference between the predictions for examining boards.

Study	Focus	Achievement assessed	Study type	Evidence weight
Delap MR (1995) Teachers' estimates of candidates' performances in public examinations	Validity	English Maths Science Sociology History Geography	Exploration of relationships	Medium

Aims
To consider the factors which influence teachers' estimates of examination performance.
To obtain some measure of how likely an estimate is to be accurate.

Study design
A sample of schools with 'A' level candidates for one examination board was asked to supply single-letter grades estimates reflecting the expected performance of the candidate in the coming examination. Multilevel analysis was used to analyse the data. The data were structured with candidate data at level 1 and school data at level 2. The function for the estimate of grade included actual grade, gender, etc.

Data collection
The subject departments in the participating institutions were asked to supply single letter estimates which reflected the expected performance of the candidates in the summer 'A' level examination. The 'A' level places candidates' performance on a scale of seven grades (A, B, C, D, E, N and U). Teachers were asked to use these same grades in their estimating.
The actual grades obtained were later collected.

Data analysis
Multi-level modelling was used to analyse the data.

Author's findings
From the distributions of estimated and actual performances, it was evident that the distributions were markedly different. 'For example, it is readily apparent that teachers were not inclined to provide estimates of low grades (N and U). Similarly more candidates obtained grade A than were estimated to do so (p 79).
The analysis of factors which influenced estimates showed the following:
(a) For most subjects, there was a significant variation between schools in the slope of the relationship between actual and estimated grades.
(b) In three subjects (biology, geography and maths), there was evidence that teachers' estimates were slightly higher for females than for males. The indications are that this pattern was followed by all of the other subjects.
(c) In the three to four months preceding the exams, the week in which the estimates were made did not substantially affect the estimated grade.
(d) The age of the candidates had a very small influence upon the estimated grade for only three of the 11 subjects.
(e) No general trend was observed concerning the estimates supplied by each type of school. The accuracy of teachers' judgements, indicated by the proportion of estimates that were accurate (obtained by the candidates) varied enormously across subjects and grade levels. For example, for physics, 84% of those estimated to gain A did so, while the proportion was only 18% for grade C; for Chemistry these figures were 28% for A and 27% for C. Overall, the proportion of accurate grades was highest for maths and biology, and lowest for physics.

Author's conclusions

'Over 7000 estimated grades were collected from approximately 450 schools in the total of 11 subjects. The analyses have shown that the teachers' estimates were not very accurate; about half of the estimates were optimistic and estimated grades of C, D or E were accurate on about one in four occasions. There is evidence to suggest that estimates for females were slightly more optimistic than those for males (reaching statistical significance in three subjects only). Potential applications of the estimated grades have been explored with the result that estimated grades can be of some value in providing information to complement decision making processes, rather than to replace them' (p 91).

Study	Focus	Achievement assessed	Study type	Evidence weight
Frederiksen and White (2004) Designing assessment for instruction and accountability: an application of validity theory to assessing scientific inquiry	Reliability	Science Practical maths/science/ tech	Evaluation: researcher-manipulated	Medium

Aims
To investigate the consistency in teachers' judgements of students' science projects.

Study design
Essentially a straightforward project focusing on the development of a scoring instrument and its evaluation.

Six middle-school teachers participated in the scoring study. Four were experienced science teachers, one was a beginning mathematics teacher, and one was a social studies teacher. Two of the science teachers were involved in the original ThinkerTools study and had considerable experience in holistic scoring of inquiry projects. The teachers participated in an iterative design process in which they tried out an initial design for the Inquiry Scorer and provided feedback about the clarity and usefulness of the scoring questions used in the project analysis and the rubrics used in the overall assessments. Using a revised version of the scoring software, they then scored 16 projects. During scoring, they scored the projects individually and then met in small groups to discuss their scoring for every fourth project scored.

Data collection
Examples of the screens used in the program for scoring are given, indicating the full set of data that the scorers were required to supply.

Data analysis
Measure of agreement among scorers was the percentage of scorers who gave the modal category of response to each question. The authors also looked at the teachers' consistency in identifying and naming the independent and dependent variables when they developed their project map. Correct coding was determined by one of the authors (JF) after reviewing each project and the scorers' responses. The consistency in rating each of the seven criteria for the overall assessments was also computed. The teachers' judgements of the overall quality were also analysed for consistency.

Authors' findings
The average rate of agreement (the percentage of scorers who gave the modal category of response to each question) across all of the project analysis questions is 81%, and the agreement rates for individual teachers range from 76% to 85%. The agreement rates vary with the number of categories available in answering a question. They are 89% for two response categories, 80% for three categories, and 75% for four categories. Thus, the teachers were, for the most part, consistent with each other in coding the local features of a project.

Teachers' consistency in identifying and naming the independent and dependent variables when they developed their project map (correct coding was determined by one of the authors (JF) after reviewing each project and the scorers' responses): Independent variables are more difficult to code (with a mean of 73% correct) than dependent variables (with a mean of 82%). In coding independent variables, the teachers had some difficulty in choosing names for them. Often they would code particular levels of a variable (e.g. 'mutts', 'hunting dogs') as though they were separate variables rather than give a name to represent the range of values or states of

the variable (e.g. 'type of dog'). The teachers remarked that developing project maps was valuable in helping them to figure out the structure of a student's project and that it made them more efficient in answering questions about the project's design and analysis.

Consistency in assigning ratings for each of seven criteria of their overall assessments (on a five point scale using a scoring rubric. The average consistency of the teachers in judging criteria was 72%, with a range for individual teachers from 63% to 79%. This is nearly as high as their average consistency in analysing specific features of projects, which was 75% (for ratings with four response categories).

Comparing the criterion ratings of the three teachers who were new to scoring inquiry projects with the two teachers who had had prior experience in holistic scoring in the ThinkerTools study: The average consistency for the new scorers, whose only experience in rating aspects of project work was in the context of carrying out a prior project analysis using the pilot version of the Inquiry Scorer, was 72%, with a range of 63% to 79%. The corresponding average consistency for the teachers with prior experience in holistic scoring was 69%, with a range of 55% to 79%. Thus, using the Inquiry Scorer with its detailed project analysis, led to a high degree of consistency among the teachers in rating important dimensions of performance related to students' goals in doing their inquiry projects, and the teachers' ability to make such judgements did not depend on their prior experience in scoring such work.

In assigning an overall project score, the consistency of the five teachers taken together was high, with a 76% rate of agreement with the modal rating. But the teachers who were new to scoring inquiry projects and learned how to score using a project analysis had a consistency of 84%, while the teachers who had had extensive prior experience scoring inquiry projects holistically had a consistency of only 62%. These results suggest that the two groups of teachers may be approaching the task of assigning an overall score differently.

Authors' conclusions
From the differences of consistency for the novice and experienced scorers, the authors concluded that the novice scorers were basing their scores on the hierarchy of judgements they made, beginning with their project analysis, followed by their evaluations using criteria related to the overall goals for doing the inquiry project. The experienced scorers, on the other hand, may have fallen back on their earlier habits of scoring by making less analytical, more holistic judgements. The systematic approach that is based on a project analysis appeared to lead to a high level of consistency. This suggested that learning to score should be regarded as a process of evidential inquiry, and should include a systematic coding of evidence and a clear mapping of evidence to criterion judgements. An implication is that, for project analysis to help teachers in making accurate ratings related to state curricular goals for students, the project analysis process needs to be designed to provide evidence that is clearly related to how student learning meets those curricular goals; conversely, the curricular goals should be stated in a way that makes them a legitimate subject of inquiry through an analysis of students' work.

The authors emphasise that establishing the credibility of judgements of performance depends on the nature of the tasks and not just on inter-scorer reliability. They argue that having open standards for tasks and how they are assessed makes possible a merging of classroom assessment goals and goals for creating evidence of students' learning that can be used by schools in meeting accountability standards. This alignment depends upon having descriptions of types of activities or tasks that provide meaningful challenges to students and teachers to aim for in learning, and also on having transparent processes for evaluating performance that can be used by teachers and students in reflecting on their work. They also argue that, in assessing students' inquiry projects, the most important issue to attend to is the internal validity of each interpretation of a students' performance, mainly that it has been properly carried out and that is accurate in characterising students' work. What is needed to accomplish this is a thorough

assessment; that is, one which develops multiple warrants for the interpretations that are made and the basis for making them.

Study	Focus	Achievement assessed	Study type	Evidence weight
Gipps *et al.* (1996) Models of teacher assessment among primary teachers in England	Validity	English Maths Science	Description	Medium

Aims
To see how teachers approached the task of conducting the required TA element of the national curriculum assessment at Key Stage 2
To find out what models they used and why they chose the approaches they did

Study design
This is a cross-sectional study, using interviews and observations across a single time point that aims to describe the ways in which teachers carry out teacher assessment, gather evidence, make decisions about National Curriculum levels and record their findings.

Twenty-nine teachers were interviewed, using a technique of 'quote sort'. This involved reading a series of quotes about assessment practices, collecting evidence, making decisions about NC levels and recording, something which had been developed through earlier interviews with Y6 teachers. The teachers were observed for a morning. In addition, five teachers who were found to be very different in their approaches were observed over four days. Qualitative analysis of the data was carried out, resulting in four clusters of teachers.

Data collection
The 'quote sort' technique was used to collect data about four aspects of the process of teachers' assessment: practices of day-to-day informal assessing; collecting evidence; making decisions about NC levels; recording.

Quotations from teachers relating to these four aspects had been collected through earlier interviews with Y6 teachers. 'The 23 quotes were shown in turn to the 29 teachers. They were asked to say whether each quote was 'very like me', ‘quite like me', ‘not really like me', not at all like me'. A recording was made of their selections and the teachers were then asked to talk about their choices - what we termed a diagnostic de-briefing interview. The only guidance given to teachers was 'tell me about why you have/have not chosen these' (p 170).

The researchers also spent a morning in each school, observing the teachers in their classes. No details were given of the observation methods, nor of the more detailed observations of five teachers who were observed for four days and interviewed in depth.

Data analysis
Constant comparison was undertaken of teachers responses in order to cluster respondents. Comparisons were made with early interviews, with observations of each teacher that had investigated their testing, assessment and recording practice, and with intensive case-study data collected in five schools. A table indicating the number of agreements between each teacher was used as a starter for the constant comparison process; indications of clusters from this constant comparison process were then checked against interview transcripts (p 170).

Authors’ findings
The four emerging models were described as teachers who were dubbed as follows:
- testers
- frequent checkers
- markers
- diagnostic trackers

For each of these the typical practice was spelled out in terms of
- ways of carrying out teacher assessment
- ways of gathering evidence
- ways of making decisions abut NC levels
- ways of recording

Testers: These do more that talking, listening and note-taking during normal activities; they plan assessment and give special tasks; they refer to levelled tasks when assigning levels; assessment is essentially 'bolt on'.

Frequent checkers: These also plan assessment tasks to be carried out at various times during the year, but also give more short informal tests of spelling and tables, more self-designed assessment tasks (aimed at groups, year groups, or sets); they also 'eavesdrop' and talk to children to pick up misunderstandings which are noted and used to inform the next day's or week's planning for teaching (but not on an individual basis); they do not like testing and data collection is an unobtrusive activity in most cases; children's performance on the small tasks or in daily activities becomes the evidence of attainment and recording a level is done more frequently than half-termly.

Markers: Their priority is teaching; they use marking schemes which later need to be converted into NC levels; work is aimed at the whole class and regular work is used rather than assessment tasks/material as evidence for assigning levels; they see marking as assessment; the work is loosely based on the NC; they are not interested in taking notes of observations and rely a lot on memory; they do not record NC attainment as they go along but convert marks from their personal marking scheme into levels for the school records half-termly or termly.

Diagnostic trackers: These are characterised by detailed planning for different NC levels, day-to-day tracking of children as they cope with the work, and TA that uses techniques of research-questioning, observation and recording incidents as they happen; they integrate assessment with teaching; they assign levels by the 'best fit' model based on the everyday work the children have done.

The numbers of teachers in each group were as follows:
- Testers: 11
- frequent checkers: 5
- markers: 8
- diagnostic trackers: 4

Authors' conclusions
Comparing the Y6 results with those from their previous work with Y2 teachers, the main themes were as follows:
- A focus on the individual and assessment for diagnosis at Y2 shifts to a focus on assessment for curriculum differentiation for the class/group at Y6.
- The strong ideological views about what is appropriate for young children shifts to a rather more accepting view of the appropriateness of formal testing by age 11.
- Along with the use of tests and assessment tasks, there would appear to be more summative than formative assessment at age 11 and this assessment tends to be less integrated with teaching.
- Informal and 'qualitative' approaches to assessment, while more evident at age 7, are nevertheless a key feature at age 11.
- At both ages (seven and 11), some teacher do not adopt the use of NC levels but rely on personal criteria.
- At both ages, some teacher collect large quantities of evidence to support their assessment.
- At both ages, some teachers are very systematic in their planning and assessment practice.

The authors are not able to make conclusions about the 'accuracy' of teachers' assessment judgements made in these different ways. They say 'it may differ, or it may not' (p 181).

Study	Focus	Achievement assessed	Study type	Evidence weight
Good (1988a) Differences in marks awarded as a result of moderation: some findings from a teacher assessed oral examination in French	Reliability	French (oral)	Exploration of relationships	High

Aims
To examine the extent to which teachers and moderators agree in their assessment of the candidates' achievements in French oral examination and to consider the possible sources of error in these assessments.

Study design
Data from pupils in six schools who were assessed in oral French at either standard or extended level. The examinations were tape-recorded and a sample marked by experienced examiners. Various statistical models for adjusting the awarded marks were applied in an attempt to arrive at a common scale across schools.

Data collection
Teachers involved were trained at a one-day training session, where the stimulus material and mark schemes were explained and discussed. Trial marking was conducted using tape-recorded interviews of pupils. Teachers assessed their pupils during the oral examinations, which were tape-recorded. Moderators assessed a sample of the tape-recorded examinations. The moderators were not aware of the teachers' marks.

Data analysis
Means, standard deviation and correlations were used to compare the teachers' and moderators' marks. There were three different methods of fitting the regression line connecting moderators' and teachers' marks, relating to three different sets of assumptions: no errors in the teachers' marks; no errors in the moderators' marks; errors in both, proportional to their variances.

Author's findings
The teachers' marks were generally more generous than the moderator's average mark. At the general level, the mean teacher's mark was 3.1 marks higher than the mean moderator's mark; this was equivalent to 0.4 grades on the oral component. At the extended level (a more open-ended interview for more advanced pupils), the mean teacher's mark was 5.3 marks higher than the mean moderator's mark; this was equivalent to 1.1 grades. The correlations indicate that there was no significant difference in variance of the teacher's and moderator's marks and the two agreed on the rank order of candidates. Most of the extended level correlations were lower than the general level correlations (i.e. taking school by school).

There were three marks allocated: the teacher's mark, the statistically adjusted teacher's mark and the moderator's mark. There were some variations, but, even in extreme cases, this was only four marks different.

Although the adjusted marks depended on the method used to fit the regression line, the difference between adjusted marks for any candidate was rarely large. The maximum difference was four marks; appreciable differences occurred only for extreme candidates. Assumption 1 (no error in the teacher's mark assumed) produced a relatively compressed mark range; assumption 2 (no error in the moderator's mark) has the opposite effects; the third assumption

has something in between. The authors suggest that the candidates should be awarded the adjusted teacher's mark.

Author's conclusions
Teachers tend to be more lenient than moderators when assessing their pupils' work; some adjustment of marks will generally be required. When given some training in the tasks to be undertaken, teachers conducting French oral examinations are able to place their candidates in a rank order that is consistent with the specified criteria nearly as effectively as assistant examiners marking conventional examination papers. However this conclusion needs to be tested in an operational GCSE context and with other subjects.

Moderators should not have access to the teacher's marks when they re-mark candidates work.

In practice, whichever version of the general statistical method is used, there will be appreciable differences only for candidates at the extremes of the achievement range in each centre. Candidates should generally be awarded the adjusted teacher's mark rather than the moderator's mark or the raw teacher's mark. There will often be considerable differences between these three marks.

Study	Focus	Achievement assessed	Study type	Evidence weight
Good and Cresswell (1988) Can teachers enter candidates appropriately for examinations involving differentiated papers?	Validity	Science History French	Exploration of relationships	Medium

Aims
To explore the issues concerning:
1. the accuracy with which teachers can predict examination performance (in order to enter candidates at the appropriate levels when differentiated papers exist)
2. the effect of the time at which the predictions are made
3. the ability of teachers to enter candidates for combinations of papers giving access to appropriate grades

Study design
This is a cross-sectional study that investigates the relationship between predicted and actual grades obtained at 'O' level and CSE and the relationship of this with the time of prediction. It also investigates the levels of entry of candidates in differentiated examinations (history, physics, French) and whether candidates had access through the papers to the grades that they then obtained. The study collects and compares actual and teacher-predicted grades of candidates entered for the GCSE examination. An experimental examination was used for part of the study and data relating to an operational CSE examination, where predictions were made at two different times, to evaluate the effect of the timing of the prediction.

Data collection
The data were lists of predictions made by teachers and actual results from the experimental exams. For a subset of the sample (those taking exams with SWEB), predicted grades for the operational CSE were also collected in May (in addition to ones provided in January).

Data analysis
Frequencies in terms of percentages of predictions corresponding with actual grades exactly, or under- or over-predicted by one, two, or three or more grades.

Authors' findings
40% of teachers' predictions were correct and a further 45% were only out by one grade. Teachers were slightly more likely to over-predict than to under-predict. Between 2% and 3% of predictions were out by more than two grades. For the CSE predicted grades, where predictions were made in January and in May before the examinations, the proportions of accurately predicted grades were almost identical on the two occasions. Thus teachers were as able to predict grades early in the year as they were later in the year.

Authors' conclusions
The authors conclude from the results of this study that, in an experimental context, teachers are able to predict the probable achievements of their pupils sufficiently accurately to enter them appropriately (from a grade standpoint) for GCSE examinations using differentiated papers. There was no evidence that being required to make entry decisions several months before the examination would create particular difficulties for the teachers.

Study	Focus	Achievement assessed	Study type	Evidence weight
Hall and Harding (2002) Level descriptions and teacher assessment in England: towards a community of assessment practice	Validity	English Maths	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 theses 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 Y2 teachers and the assessment co-ordinators in two years. Y3 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 seven year-olds and assessment co-ordinators 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, p 4): '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'.

Authors' 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 for individualistic schools); 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 pupils (contrasted with Y2 teachers working as individuals

and no grasp of the potential of enlarging the assessment community) assessment 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 initiative 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' (p 6).

'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).

Schools remained isolated from other schools in their regions in relation to assessment practice. We found that inter-school collaboration about TA and LDs was non-existent in both years; this had been a feature in all six schools in the early days of NCA.

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

Authors' conclusions
While 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 fact that funding was not made available for teachers to moderate their 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 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.

Study	Focus	Achievement assessed	Study type	Evidence weight
Hall *et al.* (1997) A study of teacher assessment at Key Stage 1	Validity	English Maths Science	Evaluation: naturally-occurring	High

Aims
To document information on the ways schools are responding to the requirements of the National Curriculum Assessment for TA at KS1
To seek 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

Study design
A sample of Y2 teachers in 45 primary schools from a single LEA somewhere in England or Wales were interviewed in June/July after completing their KS1 TA. Data were collected by (a) semi-structured interviews held, mainly at schools, lasting between 30 and 90 minutes (sometimes more than 1 teacher interviewed at a time); and (b) documentary evidence in the form of policy statements and samples of children's work and reports. Quantitative data were reported in the form of frequency of reports of use of different TA methods. Qualitative data were reported from analysis of teachers' descriptions of their approaches and perspectives

Data collection
Semi-structured interview 'to allow teachers to reveal their own attitudes to and understandings of TA and the strategies they use to assess their pupils' (p 108, Hall *et al.*., 1997). No detail of recording/ note-taking or examples of questions. No details for collection of documentary evidence.

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.

Authors' 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, though 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.

Strategies used for stage 2 are mainly observation, questioning and discussion. Only a minority

use previous records. In the ongoing process of gathering information, these same methods predominate, but there is a wide range of strategies used, including conferencing and the use of optional standard tasks and teacher-devised tasks.

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 systematic data and interpretation.
The authors report a sense of ‘professional mistrust’ amongst teachers (p 113). A wide range of assessment material was passed on from Y1 teachers but these assessments of 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 (63%) of teachers (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 pupils 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 issues 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).

Authors’ conclusions
The authors conclude that 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 raised 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 the classroom conforms to their own perceptions as revealed in this study.

Study	Focus	Achievement assessed	Study type	Evidence weight
Hargreaves *et al.* (1996) Teachers' assessments of primary children's classroom work in the creative arts	Reliability	Writing Art/music/ dance/PE	Evaluation: researcher-manipulated	High

Aims
To establish what might be called the vocabulary of primary teachers' assessment in the arts, by using a repertory grid technique (this part of the study relating to this aim is not reviewed in depth)
To establish the extent to which this vocabulary might be used in a consistent fashion by different teachers

Study design
In the part of the study reviewed here (the main study, with emphasis on the second in-service day), teachers first considered examples of pupils' work from structured and unstructured arts activities from which were drawn constructs used to develop a set of seven-point rating scales for each construct. In the second in-service day, teachers used these to scales to assess a different set of activities (presented on video-tape, audio tape and examples of children's work).

Data collection
Ratings were made against 17 scales for visual arts, 14 for music and 13 for creative writing, as listed in the paper. They rated 10 activities and their products, five of which were 'structured' and five 'unstructured'.

Data analysis
Teacher intercorrelations were explored by computing 9x9 product moment correlations matrices between the teachers' ratings of each individual product on each scale separately for each domain.

Scale intercorrelations were explored by computing product-moment correlations over all nine teachers. This was done for each of the rating scales for each of the six product categories (visual arts - structured (VA-S); visual arts unstructured, (VA-U), etc.).

Mean differences between rating of products from S and U activities were computed. These calculations were also carried out for the evaluative scales only.

Authors' findings
Correlations across scales between teachers' ratings
Mean correlations across all scales for the six categories of activities (VA-S, VA-U, M-S, etc.) were all significant at or beyond the 0.05 level. Data are given as to the proportion of individual scales where the correlation reached significance.
Mean correlations between scales were also significant at the 0.01 level or beyond. The mean correlations for structured activities in visual arts and music were greater than for the unstructured activities, but the reverse was the case in writing activities.
Ratings of products from structured (S) and unstructured (U) activities
Correlated t -tests were computed between means of the ratings of products from U and S activities over the five activities in each of the six categories. 15 out of 17 of the mean scores of the products of VA-U activities were higher than those of the products of VA-S activities, with seven of these reaching significance. When this calculation was repeated using only the evaluative scales, the difference remained, with products of U activities being rated more highly than S activities, several differences reaching significance level.

Authors' conclusions

The authors underline the 'bottom up' nature of the constructs used in the rating scales. They were based on teachers' own assessments of the products of activities which they themselves proposed (in an earlier part of the project - not reviewed here). 'The correlations between teachers across scales clearly show that there is a very high level of agreement between the nine teachers in the main study, which indicates that the 'vocabulary of assessment' can indeed be used in a consistent fashion by different individuals. This level of agreement was higher for the 'structured' than for the 'unstructured' activities in the case of the VA and M domains, which is predictable in that the former give rise to a more uniform set of products. This was not the case for the writing activities, for which the level of agreement was higher for the 'unstructured' activities (p 210). The high level of inter-correlations between scales across teachers suggests that 'teachers were applying all the scales in essentially the same way'.

'This study, although on a relatively small scale, has some important implications for future work on arts assessment. First, it has demonstrated that, when teachers are given the opportunity to clarify their ideas and the ambiguities of language used to describe children's work, they are capable of substantial agreement about the quality of different pieces of work from different pupils, and apparently make these assessment in uni-dimensional evaluative terms. Second, the more explicitly teachers define the end-product of the activity which they set, the more rigorous they seem to be in assessing the quality of this work.' (p 210)

Study	Focus	Achievement assessed	Study type	Evidence weight
Hopkins *et al.* (1985) The concurrent validity of standardised achievement tests by content area using teachers' ratings as criteria	Validity	Reading Writing Maths Science Social studies	Exploration of relationships	Medium

Aims
To investigate the concurrent validity of standardised achievement tests, using teachers' ratings and rankings of pupils' academic achievements as criteria.

Study design
Fourth- and fifth-grade students were assessed, using a battery of multiple-choice standardised tests and these results correlated with teachers' assessment. Two forms of teacher assessment were obtained and analysed: ratings and rankings within class.

Data collection
Teachers were individually interviewed by an evaluation specialist from the district office of testing. Standard interview instructions were followed in which teachers were asked to rate, on a five-point scale, their pupils' achievement in reading, mathematics, social studies, language arts, and science. Some of the teachers were also asked to rank the students in reading from best to poorest. Tests in each of these curricular areas (the Comprehensive Tests of Basis Skills, CTBS, form S, level 2) were administered about two weeks later as part of the school district's annual standardised testing programme.

Data analysis
Pearson correlations between students' raw scores on each of the five CTBS tests and the corresponding teachers' ratings (and ranking in the case of reading).

Correlation coefficients were transformed to Z-coefficients. Analysis of variance was carried out for teacher-by-content area.

Authors' findings
The average within-class correlations between the tests and the teachers' rating were 'quite high':
Language Arts: 0.74
Reading: 0.73
Maths: 0.72
Social studies: 0.64
Science: 0.60

Differences between the social studies and science correlations and those for language arts, reading and maths were significant beyond the 0.001 level.

The correlation between ratings and rankings for reading was 0.85 for the group taken as a whole.

Normalised rankings were found to correlate significantly higher than the ratings with the standardised reading tests. The superiority of normalised ranks appeared to result primarily from the reluctance of some teachers to use the full range of ratings, masking real differences between students. However, the concurrent validity coefficients of the ranks averaged only about 0.03 higher than those for the ratings.

The validity coefficients among the teachers differed greatly, although the ANOVA showed no significant teacher effect or teacher x content area interaction.

Authors' conclusions
'The high degree of correspondence between standardised achievement test score and teacher judgements, especially in language arts, reading and maths, demonstrates that both have substantial validity (or less likely, that both have little validity). Because most standardised achievement tests from the major test publishers intercorrelate highly, these results for the CBTS probably differ little from the correlations that would have been realized had a different standardised achievement test battery been used' (p 181).

They conclude that 'in general, standardised achievement tests have substantial validity' (p 182).

Although the within-class validity coefficients were slightly higher for rankings than for ratings, the differences were small. Ratings can be obtained much more quickly and with less rater frustration and appear to be a satisfactory way of obtaining teachers' assessments.

Study	Focus	Achievement assessed	Study type	Evidence weight
Koretz *et al.* (1994) The Vermont Portfolio Assessment Program: findings and implications	Reliability	Reading Maths	Evaluation: naturally-occurring	High

Aims
The formative evaluation of the Vermont Program of state-wide performance assessment.

Study design
A random sample of student portfolios of work for mathematics and writing was re-scored by second-raters. Scores on the 'uniform tests', which were part of the state-assessment system, were also collected, both in 1991-92 and 1992-93. These data were used in evaluating the reliability and validity of the portfolio assessments.

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 pupils' portfolios but in 1993 this was done by other teachers, centrally.

A sample of scored portfolios was re-scored by second raters. (There were no details of who the second raters were and how they were trained.) The 'uniform test' in writing was limited to a single prompt. The maths tests were multiple choice. No details were given of interview schedules or questionnaires.

Data analysis
Statistical methods for score data; descriptive for interview data.

Authors' findings
Rater reliability was very low in both writing and mathematics in the first year of the study. It improved appreciably for mathematics in the second year (1993), but not in writing.

The authors point out the errors that can be made by depending on percentage agreements between ratings when a scale has few values. Thus correlations are a more accurate indication.

Although the Vermont Program was not intended to provide student-level scores for external use, the study investigated these since 'the reliability of student-level scores places a bound on the quality and validity of all of the assessment results, whether individual or aggregate' (p 7). Results were analysed at three levels: (1) the score for each piece in the portfolio on each of the scoring dimensions (7 in maths and 5 in writing); (2) the dimension level -combining scores across pieces; and (3) the portfolio level.

For writing, all correlations between raters were similar at these three levels and all hovered around 0.40 for both years. In maths, piece-level correlations were low in the first year and improved in the second. Correlations at the dimension level were higher and higher still, reaching 0.79 for the eighth-grade students.

Generalisabiity showed that much of the variance in scores in both writing and maths could be attributed to disagreements among raters. In other respects, they were different. In maths, unlike in writing, the performance varied from piece to piece. This meant that 'a larger number of pieces will be needed to obtain reliable score in maths' (p 9).

In relation to validity, low reliability casts doubt on validity measures. The only other state-wide measures for comparison with the portfolio scores were the 'uniform tests' in writing (one single prompt) and maths (multiple-choice test). 'In general, the evidence pertaining to validity was not persuasive. In some respects expected relationships were found: for example, the disattenuated correlations between the writing portfolio and the writing uniform prompt were consistent with other research. But, in other instances, the relationship showed little evidence of validity: for example, the correlations between the maths portfolio score and the writing test scores were about the same as with the maths test scores.

Authors' conclusions
The unreliability of the scoring alone was sufficient to preclude most of the intended uses of the scores. Efforts to examine the validity beyond the reliability of scoring were hindered by both conceptual and empirical obstacles, but preliminary analyses showed ground for concern (p 7).

'Our observations of the Vermont scoring also suggested that the Vermont rubrics and the training of writing raters may have played a role. During bench-marking sessions, we observed considerable confusion among writing raters about the interpretation and application of the rubrics and raters sometime disagreed about which scoring dimension was germane' (p 12). Two factors contributing to the unreliability were the difficulty of training a large number of raters and the lack of standardisation of tasks. This lack of standardisation required raters to stretch general-purpose rubrics to cover a wide variety of tasks. By contrast, many performance-assessment programmes that have demonstrated high levels of rater reliability have relied on standardised tasks and have used task-specific rubrics. An intermediate approach would be to allow unstandardised tasks but to apply genre-specific scoring rubrics (see Gentile, 1992).

The problem of validity in the Vermont experience is likely to arise elsewhere. The problem is that portfolios do not provide reasonable samples of the domain and the sampling of tasks may vary greatly from one classroom to another. Finally, '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 compromises between them' (p 15).

Study	Focus	Achievement assessed	Study type	Evidence weight
Levine *et al.* (1987) The accuracy of teacher judgement of the oral proficiency of high school foreign language students	Reliability	French and Spanish oral proficiency	Exploration of relationships	Medium

Aims
To determine the accuracy of teacher judgement of the oral proficiency of pupils in French and Spanish high school classes.

Study design
Data from pupils in six schools were assessed in oral French at either standard or extended level. The examinations were tape-recorded and a sample marked by experienced examiners. Various statistical models for adjusting the awarded marks were applied in an attempt to arrive at a common scale across schools.

Data collection
Using a table of random numbers, four student-participants and three alternates from each of eight teachers' classes were randomly selected to be rated on the American Council on the Teaching of Foreign Languages (ACTFL) oral interview. Teachers were shown copies of the ACTFL scale and then asked to predict where each pupil would fall. Finally, teachers provided person information and information about the selected pupils. The following week, each language group of pupils (French and Spanish) was rated in an oral interview by an independent tester. Both testers were certified by ACTFL.

Data analysis
Direct comparison of teacher predicted and actual rating, analysis of variance to explore difference across sub-groups, analysis of variance in relation to student and teacher data.

Authors' findings
There was a significant difference between teachers' predicted ACTFL rating (mean 4.9) and students' actual rating (mean 3.4). The teachers consistently overestimated their students' ability. The means for the French and Spanish groups were not significantly different.
The number of years of instruction in the foreign language did not influence the difference between estimated and actual ratings. Regardless of the number of years that a student had spent studying, teachers consistently over-rated their oral performance.

In relation to letter grade, there was a definite trend. 'A' students were overestimated by a greater amount than 'B' students who in turn were overestimated more than "C" students.
In relation to groups based on the actual ACTFL score, teachers overestimated more for students performing at the lower end of the scale. The interaction between actual rating and letter grade was approaching significance. The authors suggested that this means that the academic letter grade severely biases teacher judgements.

Authors' conclusions
The findings that teachers consistently overestimate pupil oral performance and that the academic letter grade influences teachers' judgement led the authors to suggest that an oral proficiency workshop is justified. In such a workshop, the consistent overestimation and letter grade influence would be directly addressed. They cite evidence that training programmes can make teachers more accurate in their judgements. (Liskin-Gasparo reports remarkable consensus after only an hour's training.) They claim that 'teacher themselves would be receptive to assessment training' (p 50). 'Perhaps the most important by-product of assessment training

may be that teachers will begin to modify their curriculum to make day-to-day classroom activities more closely congruent with the increased emphasis on oral language proficiency' (p 50).

Study	Focus	Achievement assessed	Study type	Evidence weight
Meisels *et al.* (2001) Trusting teachers' judgements: a validity study of curriculum-embedded performance assessment in kindergarten to Grade 3	Validity	Reading Writing Maths	Exploration of relationships	High

Aims
To examine the relationship of curriculum-embedded performance assessment to other key indicators of student achievement
To investigate the validity of the Work Sampling System (WSS), a performance assessment for pre-school to Grade 5, by determining whether teacher judgements about student learning are trustworthy when those judgements are based on the WSS

Study design
The study was carried out in 1996 in seven schools where WSS had been implemented for three years. The WSS was used by 17 teachers of pupils Kg to Grade 3 throughout one year. This involved using the checklists on three occasions, collecting portfolio material at all times and creating summary reports on three occasions. These results were compared with the result of the Woodcock-Johnson Psycho-educational Battery - Revised (WJ-R) on two occasions during the year.

Data collection
The WSS checklist comprise skills and behaviours presented in the form of a one-sentence performance indicator. Teachers rate students' performance on each item of the checklist three times per year and compare the rating with national standards for children of the same grade. The items in the checklists (differing for each grade) measure seven domains of development: personal and social; language and literacy; mathematical thinking; scientific thinking; social studies; the arts; and physical development. Teachers use a modified mastery scale: 1= not yet; 2 = in process; 3 = proficient. During these periods, the teacher also completes the summary report, summarising each child's performance in the seven domains and rating it as 1 = 'as expected', or 2 = 'other than expected' and compare it with past performance. Portfolios collect work that illustrates students' achievements, efforts and progress.

The WJ-R battery was administered individually to each child by examiners trained in its use on two occasions during the year.

Data analysis
Subscale scores for the checklist were created by computing the mean score for all items within a particular domain (i.e. language and literacy or mathematical thinking). Subscale scores for the summary report were created by computing a mean for a combination of three scores: students' checklists and portfolio performance ratings, and rating of student progress. When computing the subscale scores, missing data in the teachers' WSS rating were addressed by using mean scores instead of summing teachers' ratings.

Three analyses were conducted with the cross-sectional data using:
(a) correlations between students' standard scores on the sub-tests of the WJ-R and the WSS checklist and summary report ratings of student achievement within the corresponding WSS domains
(b) four-step hierarchical regressions to examine the factors that accounted for the variance in students' spring WJ-R scores

(c) Receiver-Operating-Characteristics (ROC) curves to determine if a random pair of average and below average scores on the WJ-R would be ranked correctly in terms of performance on the WSS.

Authors' findings
Three-quarters of the correlations between WJ-R and WSS were within the range 0.50 to 0.75 and 48 of the 52 correlations between the WSS and the comprehensive scores of children's achievements fall within the moderate to high range. On the authors' judgement that correlations of 0.70 to 0.75 are optimal, these findings are taken to indicate 'strong primafacie evidence for the concurrent aspects of WSS's validity' (p 84).

The four-step regression analyses indicate that significant associations between WSS spring ratings and WJ-R spring outcomes remained even after controlling for the potential effects of age, SES, ethnicity, and students' initial performance level on the WJ-R in literacy (Kg 2) and in maths (Kg 1).

There were some differences across grades. In the first step of the regressions, only the demographic variables were entered. This model was significant only in Kg and second grade for language and literacy and in Kg for maths. When entered into the second step of the regressions with the demographic variables, the WSS checklist was significant at all grades levels for both maths and literacy. It explained more than half of the variance in literacy scores in grades 1 and 3. When the summary report was entered in the third step, both the summary report and the checklists contributed significantly in explaining the variance in the spring WJ-R literacy score for Kg 2. In the third grade, the checklist alone was a significant predictor of the language and literacy score. In maths, the WSS variables were significant predictors in step 3 of the regressions for Kg 3.

Authors' conclusions
'The results of the correlational analyses provided evidence for aspects of the validity of the WSS. The WSS demonstrates overlap with the standardised criterion measure and makes a unique contribution to the measurement of students' achievement beyond that captured by WJ-R test scores. The majority of the correlations between the WSS and the comprehensive scores of children's achievements (broad reading, broad writing, language and literacy and broad maths) are similar to correlations between the WJ-R and other standardised tests.(Correlations between the WJ-R and other reading measures of 0.63 to 0.86)' (p 89).

Analysis of the mean WSS scores in the third grade indicates that teachers overestimated student ability on the summary report compared with the WJ-R.

'Overall the regression results provide evidence that WSS ratings demonstrate strong evidence for concurrent aspects of validity, especially regarding students' literacy achievement' (p 90).

'The findings of this study demonstrate the accuracy of the WSS when compared with a standardised, individually administered psycho-educational battery....and it is a dependable predictor of achievement ratings in Kg to Grade 3. ...it discriminates accurately between children who are/are not at risk' (p 91)

Study	Focus	Achievement assessed	Study type	Evidence weight
Papas and Psacharopoulos (1993) Student selection for higher education: the relationship between internal and external marks	Validity	Science Law Economics	Exploration of relationships	Low

Aims
To investigate, in the context of education in Greece, the correlation between internal marks of candidates applying for higher education and the marks gained on external examination marks. Until 1988, a combination of these two kinds of marks was used for university entrance decisions. The hypothesis is that, if these are highly correlated, then there may no need for external examinations.

Study design
The correlation of the results of two methods of assessment for a sample of higher education applicants from different schools.

Data collection
It is not clear as to whether the data were collected directly from the students (a student questionnaire is mentioned) or from the schools.

Data analysis
Calculation of means and standard deviation of internal and external marks; correlations between internal and external marks and regression, used to predict external marks, controlling for internal marks, school type and subject cluster.

Authors' findings
In terms of mean internal and external marks, there is a substantial difference by school type and subject cluster. Private schools and selective schools are closer in match between external and internal marks, and for those aspiring to enter medical or law school. However, for the correlations, 'non-selective schools report higher correlations than selective schools, with the exception of the polytechnic cluster' (p 399). Public schools report higher correlations than private schools.

'But all correlations are on the high side revealing that the external marks somehow validate internal school marks' (p 400).

Authors' conclusions
They raise the question as to whether external examinations are really necessary. 'The marks in the last three grades of secondary school seem to reflect fairly well how a student will perform in the external examination' (p 401).

Study	Focus	Achievement assessed	Study type	Evidence weight
Radnor H (1995) Evaluation of Key Stage 3 Assessment in 1995 and 1996. Evaluation of Key Stage 3 assessment arrangements for 1995	Validity	English Maths Science	Evaluation: naturally-occurring	Medium

Aims
To evaluate the assessment arrangements in 1995 and 1996 at KS3 and to report issues associated with the implementation of the statutory tests in the subjects of English, maths and science and teacher assessment from the perspective of the schools. Data are extracted from only one part of this study, that concerning teacher assessment. However, reference is made here to some data relevant to the validity of the national curriculum tests.

Study design
39 schools constituted a 'core group', which were visited on two occasions. The core group provided pupils' scripts, which were scrutinised. Visits to the core group informed the development of questionnaires which were sent to a larger number of schools spread across the regions of England and Wales.

Data collection
Visits to 39 schools; questionnaires sent to 317 schools and the scrutiny of about 2000 pupils test scripts, all relating to the Key Stage 3 arrangements for 1995 and 1996.

Data analysis
Simple descriptive statistics are given for questionnaire responses in terms of percentage response. No information is given about how data from visits was used or how the scrutiny of tests was carried out.

Author's findings
In relation to the quality and relevance of tests in English, the study found that in one paper there was evidence of pupils not understanding the expectations of the questions. There were also multiple interpretations of the mark schemes. The Performance Criteria allowed markers to make professional judgements and to award marks on a best-fit basis. However, it was noted that the shift from providing generic statements describing the nature of responses in 1995 to providing tentative statements which describe possible responses (involving the words 'might' and 'may') in 1996, did not assist the process of assessment (p 49). Markers were reluctant to use the full range of marks. The external marking of English elicited more negative than positive views.

In mathematics, about half the teachers interviewed felt that the tests were as much a test of English as of mathematics, with the wordiness of the questions disadvantaging some children. There was evidence in the scripts that certain contexts forced errors in students, but there were few marking errors. In science, however, tight marking schemes did not include some correct responses.

In order to complete their teachers' assessment, English teachers found written work completed in class most effective for gathering evidence. Maths and science teachers tended to rely on school examinations and tests. All teachers believed cross-moderation among teachers to be important, but the constraints of finance and time made this difficult. The tendency was, instead, for individual teachers to use standardised test material or work with standardised exemplar materials, such as those provided by SCAA/ACAC.

As far as the relationship between TA and test results was concerned, teachers were divided into 'levellers' and 'differentialists'. Levellers expected the two to show comparable levels and they finalised their TA after the test results were considered. Differentialists did not expect a match, considering that TA and tests assess different things. They completed their TA without taking tests results into consideration. English teachers were mostly differentialists, while levellers predominated in the maths and science teachers.

The majority of teachers made little use of scripts returned to the school after marking.

Author's conclusions
These are the same as the findings.

Study	Focus	Achievement assessed	Study type	Evidence weight
Reeves *et al.* (2001) The relationship between teachers assessments and pupil attainments in standards test tasks at Key Stage 2, 1996-98	Reliability	Reading Writing Maths Science	Exploration of relationships	High

Aims
To explore the relationships between pupil attainments on standard National Curriculum tests at the end of KS 2, teacher assessment and pupil characteristics of gender, age, EAL and special needs, using representative samples drawn from school in England in 1996, 1997 and 1998.

Study design
A cross-sectional study of the relationship between the two measures of pupil attainment provided by TA and national tests for a national sample of pupils and the relationship between any difference between these measures and pupil variables.

Data collection
Teachers based their assessment on classroom observations and classwork, and were able to take the national tests into consideration if they wished. The extent to which this occurred is not known.

The data were provided by the schools as part of the wider project of which this study was a part, marked scripts were also provided by the schools and returned after scrutiny by the researchers.

Data analysis
ANOVA was chosen in preference to multilevel modelling on account of the low numbers of pupils per school.

Simple percentages of pupils at each level for each subject for each year, percentage agreement between TA and national tests for each level for each subject for each year.

Authors' findings
Impact of school and pupil characteristics on attainment: School had a considerable effect on both TA and tests results, explaining between 17 and 26% of the variance in attainment levels.
Age: Older pupils achieved higher levels in all subjects.
Sex: Males outscored females in maths and science, but the reverse was true in English.
SEN: This had a high impact in all cases. Together SEN and school accounted for 40% or more of the total variance.
EAL: This had a significant effect in 1998 (when the sample was increased), with pupils whose first language was not English receiving lower levels than average.

Comparison of test results and TA: 'Comparisons reveal a remarkably high level of consistency across years in all three subjects' (p153). But there were differences in the direction of the differences across subjects. For mathematics, test levels were lower than TA (significant for 1996 and 1998 but not 1997). In English and science, the opposite was found. However, although significant, the differences by subject were small.

ANOVA was used to explore the relationship between pupil characteristics and the size of the difference between TA and test scores. It was found that the school had a big impact. The amount of variance explained by this factor declined over time in all subjects, most notably

science. Age had only a small effect. Sex had a significant impact for mathematics across all years, with teachers consistently under-rating boys more than girls. The same was found for science, but for English the opposite was found: girls were more frequently under-rated. SEN was significant for English and science in all years and for maths in 1998 only. In all cases, TA levels were more likely to be lower than test results where pupils had SEN. In many instances, the effect was considerable: for example, 24% of pupils with SEN were awarded lower levels than the test results for science in 1998. The only effects of EAL were in English in 1998. Teacher under-rated 25% of EAL pupils compared with 15% of others.

Authors' conclusions
From the reduction in the schools effect over time, it was concluded that schools have become more similar to one another in terms of patterns of differences between test and TA results. However, agreement between test results and TAs have remained remarkably consistent over time. Considering the degree of relationship between TA and levels on standard KS tests, a certain amount of non-agreement is not only acceptable but even desirable, otherwise one measure or other becomes redundant.

The agreement remained consistently across the years at about 75% and less than 0.5% of disagreements exceeded one level on the rating scales. 'This finding would seem to fit the bill of having two complementary assessment measures which usually concur but which vary sufficiently to justify maintaining the use of both' (p 158).

Subject, sex, age and EAL has some varying effects on the difference between TA and test scores. However, the strongest relationship was evident for SEN, with SEN pupils tests frequently exceeding their teachers' assessment levels. This happened in all subjects, but particularly in English and science. The authors offer alternative explanations to the obvious one for this: that teachers 'teach to the test' for these pupils, who therefore may be able to perform in the tests above their real attainment level.

Study	Focus	Achievement assessed	Study type	Evidence weight
Rowe and Hill (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles'	Reliability	Reading Writing Maths	Evaluation: naturally-occurring	Medium

Aims
To obtain data from the use of the 'subject profiles' in order to report on reliability and validity of teachers' assessment using these profiles and to provide examples of their use in plotting pupils' progress.

Study design
A cross-sectional study in which ratings by teachers were collected within the context of two projects and used to provide evidence of reliability and validity of the rating scale (subject profile).

Data collection
Teachers were asked to rate a student's level of achievements with reference to the indicators for each of the nine bands (A to I) of the reading, writing and spoken language strands of the English profiles, and for each of the twelve levels (1 to 12) of the number and space strands of the mathematics profiles. These were recorded on class recording lists, where a score of 3 is typically recorded, if all the behaviours associated with a given band/level are consistently displayed by the student; 2, if most of the behaviours are present; 1, if some of the behaviours are beginning to be developed; and 0, if none of the behaviours have yet been observed. The ratings for each band/level are then added together to give a total score out of 27 for each English profile strand, or 36 for either strand of the mathematics profile (p 327).

Data analysis
The Guttman methods of scaling (Guttman standardised item alpha coefficient), test-retest reliabilities and inter-rated reliabilities were calculated.

Authors' findings
The Guttman reliability estimates indicate that the profiles appear to function as cumulative scales or growth continua and that teachers are consistent in their use of them. Coefficients ranged from 0.77 to 0.96.

'The reliability coefficients were not as high for early years as for later years, on account on the restricted range in the achievement levels of students in earlier years. In addition, with the exception of the preparatory grade (Kg), estimates for the two mathematics strands are somewhat higher than for the three English strands' (p 128).

Pearson product-moment correlations between teacher assessments of the same students made on two occasions four months apart, indicate high test-retest reliability (r values from 0.89 to 0.93). Inter-rater reliability (different teachers, same students) were also high (0.83 to 0.89).

Authors' conclusions
The evidence suggests that, when using subject profiles, teachers are consistent in their assessment of students and are also able to achieve a satisfactory level of inter-rater reliability (although the current evidence is only partial on this point). It is also relevant to note that, among teaching staff participating in the VQSP (Victorian Quality Schools Project), there is agreement that assessments based on Victorian subject profiles have a high degree of validity in terms of measuring students' levels of achievement in English and mathematics as taught in their

schools, with such views being most frequently expressed for elementary school English and least frequently for secondary school maths.

Study	Focus	Achievement assessed	Study type	Evidence weight
Shapley and Bush (1999) Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience	Reliability	Reading Writing Maths	Evaluation: naturally-occurring	High

Aims
To examine the extent to which the portfolio system introduced in Dallas, Texas, met appropriate technical standards for its intended uses.

Study design
An evaluation in which data from the scoring of portfolios by students' own teachers was compared with scoring by other trained teachers and with scores on a standardised test battery of basic skills. Second raters also judged the adequacy of the evidence in the portfolios for making assessment decisions.

Data collection
Portfolio scores: 'The Reading/Language Arts Portfolio Assessment is a comprehensive assessment system designed for students in pre-Kg through second grade that included four interrelated elements:(a) student work samples, (b) instructional goals and performance criteria, (c) student summary and (d) scoring rubrics' (p 115). Student work sample are accumulated over the year by the class teachers, according to a set of guidelines. The work samples are to align with goals and performance criteria on the Texas content standards. A minimum of 12 student work samples is required.

Pre-Kg portfolios are rated on four goals and grades 1 and 2 on five goals. The scoring rubrics align with the instructional goals. Scoring is on a four-point scale: from 1 = emerging to 4 = distinguished. In addition to scoring individual samples of work, teachers make an overall judgement of how well a collection of work samples meets the multi-dimensional standards that define each instructional goal.

Second raters had extra training and the teachers' ratings were removed before they rated the students' performance.

Second raters recorded information in two ways. First, instructional goal ratings were made on a student summary form for the spring rating period. They noted if the evidence was not sufficient to allow rating. Second, they completed a form recording the adequacy of documentation, such as the presence of details (e.g. grade level, goal and performance criteria).

Scores on the Iowa Test of Basic Skills (ITBS) were also collected for the relevant pupils; it is assumed that these were collected as routine.

Data analysis
Mean score differences between teachers' and second raters, percentages of agreement and interrater correlations, correlations between portfolio scores and ITBS raw scores.

Authors' findings
Mean scores for the four instructional goals recorded at Kg and pre-Kg: All teachers' ratings were higher than second raters' scores, reaching significance in all cases except for 'writing about experiences'.

For grades 1 and 2, all teachers' ratings were higher than second raters' scores, reaching significance in all cases except for listening and speaking.

In terms of consistency, the percentage agreement between teachers and second raters varied by instructional goal and grade level. 50% to 59% of grade 1 and 2 portfolios received exactly the same score by two raters when evidence was adequate. When raters were not in perfect agreement, a difference of one point was most likely. However ‘ratings could not be made about half of the time for many instructional goals because of inadequate evidence' (p 12).

Interrater correlations were low. A mean pre-Kg and Kg correlation of 0.37 indicated that 14% of the variance in second raters' scores was explained. Likewise, the first and second grade mean correlation of 0.48 indicated that 23% of the variance in second raters' scores could be predicted from knowing the teachers' scores. Large percentages of unexplained variance were due to error.

Validity based on second raters' analysis was low. ‘Overall, for many portfolios, there was insufficient sampling of the content domain because the number of samples was inadequate, the work sample provided inadequate information, the purpose for the work samples was unknown, the work samples did not exemplify the content knowledge of the goals, or there were no teacher notations to explain activities and to clarify the student performance' (p 123).

For convergent validity, there were low positive correlations between the portfolio ratings and closely related ITBS sub-test scores. Divergent relations between ratings and mathematics scores were not firmly established. 'The differences between the math sub-test score associations pointed to confounding effects of reading and language development when mathematics tests required students to read and solve written problems and to interpret data' (p 126).

Authors’ conclusions

'After three years of development, the portfolio assessment did not provide high quality information about student achievements for either instructional or informational purposes. The unreliability of the scores was likely related to (a) lack of standardisation of tasks, (b) problems with the scoring rubrics and (c) inadequate training....Because the tasks were unstandardised, the scoring rubrics aligned with the instructional goals rather than with specific tasks....Improving the scoring rubrics will require greater standardisation of the portfolio contents so that there is stronger alignment between the tasks and specific evaluative criteria. This suggests a need for a compromise between standardisation, which is needed to improve technical quality, and the flexibility that allows portfolios to be integrated with the classroom context' (p 127).

The authors suggest that it seems that portfolios need to contain a core of essential work samples (those that all portfolios much contain) and optional work samples (those that the teacher and students agree to select). The core samples would provide a common frame of reference across all portfolios for judging students' performance. 'Altogether, the findings of this study raise cautions to those who believe portfolios are the answer to the perceived ills of standardised testing for young children' (p 128).

Study	Focus	Achievement assessed	Study type	Evidence weight
Sharpley and Edgar (1986) Teachers' ratings vs standardised tests: an empirical investigation of agreement between two indices of achievement	Validity	Reading Maths General attitude Verbal intelligence	Exploration of relationships	Medium

Aims
To study the accuracy of teachers' ratings of reading vocabulary, reading comprehension, mathematics, and verbal intelligence when compared with standardised test scores and to explore if teachers' attitudes towards students biased their evaluations of students' progress and to explore if bias against boys exists in teachers' evaluations.

Study design
The study compares the teachers' ratings of their pupils with scores on standardised tests of the same aspects of achievement.

Data collection
For the teachers' ratings: A standard rating form with a table of ratings from 1 to 5 (1 = high, 5 = low) on which teachers rated each child in their classes according to their present level of achievement in the four areas and general attitude.

For the Progressive Achievement Tests (PAT), pupils were tested in groups of 20 to 30 and were required to record their answers on the PAT mathematics answer sheet and the ACER Standard A100 OMR answer sheet. The PAT sessions were all conducted by the same post-graduate student in special education. For the Peabody Picture Vocabulary test, there was individual administration by post-graduate students in psychology who were trained and judge competent in administration by a state-accredited psychologist.

Data analysis
Statistical methods for comparison of scores and ratings.

Authors' findings
The teachers' ratings for girls, on the five-point scale, showed higher ratings for general attitude and verbal intelligence than for reading and maths. For boys, verbal intelligence was rated highest and general attitude lowest of the five ratings.

Ratings on comprehension and vocabulary were significantly associated as were comprehension and maths. General attitude was not significantly correlated with ratings on the other variables.

The standardised tests and the teachers rating showed girls scoring higher that boys on vocabulary, reading comprehension and mathematics, but not on verbal intelligence, where differences were negligible.

Correlations between ratings and test scores were significant at the 0.01 level in 19 out of the 20 correlations for boys and 15 out of the 20 correlations for girls. However, although significant, they were low and even the highest (0.56 for girls comprehension scores and ratings) accounted for only 31% of the variance.

Authors' conclusions
Given that the aim was to examine the accuracy of teachers' ratings when compared with standardised test results, the correlations do not support a strong commonality between the two assessment procedures. 'It appears from the data that, although there are significant correlations between teacher ratings and standardised test scores, there is little direct meaning in these results'. The authors suggest that teachers' ratings and test scores assessment constitute two 'non-equivalent domains'.

Teachers were no more accurate in assessing boys or girls. The results for verbal intelligence, showing overlap between teachers' ratings and PPVT-R only to the extent of 2% of the variance (for girls and 15% for boys) do not provide evidence that teachers can accurately assess a child's verbal intelligence. 'What we suspect has been indicated by these data is the development of two (nearly) parallel domains of achievement and two sets of assessment criteria that require attention to develop communality' (p 110).

Study	Focus	Achievement assessed	Study type	Evidence weight
Shavelson *et al.* (1992) Performance assessments: political rhetoric and measurement reality	Reliability	Science	Evaluation: researcher-manipulated	Low

Aims
To evaluate the reliability and validity of science performance assessment.

Study design
A cross sectional study in which a group of students were given a range assessments, so that the results could be compared and correlated.

Data collection
The performance assessments were developed in the following forms:Three hands-on investigations ('paper towels', 'electric mysteries', and 'bugs') were created. When conducting these investigations, the students were observed and they also used notebooks to record specific aspects of the investigations; these were collected as a second mode of assessment. Computer simulations were created for two of the investigations (omitting 'paper towels'), then short-answer and multiple-choice questions were chosen to parallel in content the three hands-on investigations.

The students undertook these performance assessments in the following order, separated by three weeks: paper-and pencil measures, observed investigations with notebooks, and finally computer simulations. The students were from two school districts, one being noted for its hands-on science curriculum and the other having no regular science curriculum, apart from a textbook course on health.

Data analysis
Frequencies of scores for the two school districts, means performance levels, correlations between observers, correlations between the various assessment modes and between the aptitude test scores and the standardised science test and between aptitude scores and the 'benchmark' hands-on investigations.

Authors' findings
For the observed investigations, the following was found:

- Interrater reliability was consistently high for all investigations and varied little by student' curricular experience.
- Inter-task reliability (internal consistency) was difficult to attain. Some students performed well on one task but poorly on another.
- For all investigations, mean performance was higher for students from the 'hands-on' science district than for the 'textbook' science district.
- The correlations between investigations and standardised multiple-choice tests were only moderate in magnitude, suggesting that these tests measured different aspects of science achievement.
- The correlations between aptitude and the investigations were lower than between aptitude and the standardised science test (This difference was as hypothesized.)

For the other forms of test:

- Notebooks provided the closest approximation in reliability and validity. The next closest surrogate for observed investigations was the computer simulations. Mean performance was comparable to the 'benchmarks', as were the patterns of correlations. However, some

students who scored high on the benchmarks scored low on the computer simulations and vice versa.

- The paper and pencil surrogates did not fare as well. Compared with the benchmark (observed investigations, the short-answer items were less reliable and correlations with the standardised achievement test and aptitude test were higher. Moreover, mean performance of students experienced in hand-on science did not differ significantly from the performance of the students receiving 'textbook' science.
- Multiple-choice items fared even worse.

Authors' conclusions

- Raters are able to reliably evaluate student hands-on performance on complex tasks in real time. Reliabilities are high enough that a single rater can provide a reliable score.
- Task-sampling variability is considerable.
- Performance assessment can distinguish between students with different instructional histories. Assessments that are closely linked to a specific domain of knowledge (e.g. electric circuits) are more sensitive than more general process assessment (e.g. paper towels).
- Performance assessment must be carefully crafted to measure more than science aptitude. To be curriculum-sensitive, they need to measure the application of both declarative and procedural science knowledge.
- Notebooks and computer simulations can serve as surrogates for actual investigations.
- There is considerable variability among students in relation to the particular investigation and to the particular method used for assessment.

Study	Focus	Achievement assessed	Study type	Evidence weight
Shorrocks *et al.* (1993) Testing and assessing 6 and seven year-olds. Evaluation of the 1992 Key Stage 1 National Curriculum Assessment Final Report	Reliability	Reading Writing English Maths Science Practical maths/science/ tech	Evaluation: naturally-occurring	High

Aims
To 'report the views of a representative sample of teachers carrying out the 1992 tests and to provide an academically rigorous evaluation of the results of those tests' (p 2). (Note that only the latter aim is represented in the data-extraction.)

Study design
A cross-sectional study of a national sample of pupils in which pupils' TA and SAT scores were compared and variations with pupils background characteristics were investigated for tests in English, maths, science and technology. Information was also collected from teachers' logs and questionnaires, but not used in this data-extraction. A review of the impact of gender, ethnicity and age was also included in the study.

Data collection
While questionnaires are not given, a source is indicated. TA and SAT details are not given, these followed the requirements of the national curriculum assessment procedures.

Data analysis
Statistical methods were used to analyse TA and SAT scores to given distributions across levels, and to compare these distributions. The performance profiles of groups formed by different pupils characteristics (sex, age, etc.) were also presented.

Authors' findings
In terms of the distribution across levels, there were variations across the attainment targets within English, maths and science. There were identical distributions for TA and standard tasks for two of the attainment targets in English, but large differences for aspects of mathematics and science. The Kappa statistic of agreement comparing actual differences with agreement by chance showed that, in many cases, agreement between the two sets of scores was high. There was close agreement for English but less in some aspects of maths and science. The authors caution that a proportion of individual children may be placed in different levels by the two assessments, but the overall numbers at each level may be the same. The levels of agreement were considerably greater than those found by the same authors in 1991.

Performance of different groups of children was reported as follows:
Gender: TAs suggested statistically significant differences in favour of girls for English and one aspect of maths (using and applying). Standard task analysis showed the same pattern.
Age: Winter-born children (i.e. older) had statistically higher scores at the subject level for English, maths and science.
Social background: At subject level, there were statistically significant differences in favour of children from higher neighbourhood status areas.
Ethnic groups: Significant differences were found, but these varied and only in English did the white children appear to be superior.
EAL: At subject level, there were significant differences between groupings in maths and

science. Differences in favour of English speaking children were found in some aspects of all subjects.
SEN: In all subjects, there were significant differences, in favour of those without SEN in both TA and standard tasks.
Class size: Here the results were contrary to expectations (given that small classes are found more frequently in low socio-economic areas), showing a positive effect of small classes.

Authors' conclusions
By comparison with 1991 results, performance levels were higher. The authors suggest several reasons for this: greater teacher confidence in the assessment procedures, the publication of results, which raised the stakes (results in 1991 were not published) or greater teaching to the test.

In relation to variations for different groups of children - when schools are being judged by the outcomes of the assessment, 'younger children in a year group, children from poor social backgrounds, children for whom English is a second language, and children with special needs seem to be a distinct liability on a school roll' (p 57).

Study	Focus	Achievement assessed	Study type	Evidence weight
Thomas *et al.* (1998) Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment	Reliability	Reading Writing Maths Science	Exploration of relationships	High

Aims
To study the operation of the national Curriculum Assessment of seven year-olds in England to address equity issues and to compare the outcomes with score on a standardised word recognition test.

Study design
A cross-sectional study in which pupils are all administered the same assessments and variations in performance are explored in relation to pupil characteristics.

Data collection
Data collected by researchers from one LEA. The data on pupils were collected by teachers: TA by teachers' observation of pupils during regular work and by teachers administering standard tasks to individual pupils following instructions.

In both TA and ST assessments pupils are assessed as being at one of five levels (working towards level 1, or at levels 1, 2, 3 or 4.

No information is given about how the NFER word-recognition test was administered.

Data analysis
Comparisons of TA and ST levels, examination of relationship between the five characteristics of pupils and TA, standard tasks and the standardised word recognition test by means of multilevel modelling.

Authors' findings
Relationship between TA and standard tasks
There was a fairly strong positive relationship overall between equivalent standard task and TA scores, with correlations ranging from 0.92 for reading and 0.77 for maths probabilities score.

Correlations between TA, standard tasks and student characteristics
There was a statistically significant relationship between nearly all assessment scores (TA and standard tasks) and the five student intake characteristics, irrespective of the method of assessment. Across all subjects, three factors – frequency of free school meals, EAL and SEN – have a stronger impact on attainment than the other two background factors (gender and age). 'The multilevel model results indicate that, for the most part, there are statistically significant differences between the average performance of different student groups categorised by gender, income, language, special needs and age groups across subject levels, TAs, standard tasks and the word recognition test' (p 222).

The impact of pupils characteristics
Girls perform at a higher level than boys on English and mathematics (no sex difference in science) and, in all three subjects, those entitled to free school meals performed at a lower level that those not entitled. Students with SEN perform at a substantially lower level than those without, and older pupils perform better than younger on average. The school factors of mean age and percentage free school meals both have a negative impact on attainment across the

three subjects (i.e. there is slight tendency to lower performance, on average, the older the mean age of the Year 2 pupils within a school).

The relationship between standard tasks and TA
'The evidence suggests that across all subjects and aspects of subjects, teachers are systematically assessing students differently on the TA in comparison to the equivalent standard task and that schools (and teacher judgements) vary in the way TA are scored, even after standard task results are taken into account. The results show that students' standard task assessments account for between 59% and 94% of the variation between schools in the TA. In most cases, once the ST levels have been accounted for, each student background characteristic still has a small but statistically significant impact on TA level' (p 223).

Authors' conclusions
For the most part, TA and standard task assessment worked similarly. Nonetheless, in some instances, TA results were more likely to widen the gap between groups of students - at least modestly - especially for those who do and do not have a statement of SEN (p 228).

The results of the word-recognition test show the lowest percentages of variation in student scores attributable to schools (both before and after controlling for pupil background) suggesting, as might be expected, that this test is less susceptible to a lack of equal standards in the scoring criteria than the NC reading assessments. These findings suggest that, as well as considering the mean attainment differences between particular groups of students on 'authentic' assessments, it is also vital to consider the possibility that different teachers may interpret the assessment criteria differently and that the importance of this factor may vary according to the assessment domain.

The topic of teacher judgement has also been addressed by examining the variation across schools in TA after controlling for each student's standard task attainment as well as their background characteristics. The findings indicate that students' standard task scores explain the majority of the variation in teacher assessed scores but, of the remaining variation, there remains a substantial proportion attributable to schools. The impact is greater for some subjects than others (p 235).

Study	Focus	Achievement assessed	Study type	Evidence weight
Wilson and Wright (1993) The predictive validity of student self-evaluations, teachers; assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachia schools	Validity	Maths Other Verbal reasoning	Exploration of relationships	Medium

Aims
To examine a) the accuracy of student self-estimates of academic performance, b) the degree of relationship among estimates of performance provided by students, teacher assessments of ability and success, course grades and scores on a standardised measure of verbal ability and of numerical ability and c) the degree to which student self-estimates, teacher expectations and grades were predictive of performance.

Study design
A cross-sectional study of the relationships among a number of measures of students' verbal reasoning and numerical ability in order to estimate how well various students and teacher estimates of performance predicted actual performance on standardised tests.

Data collection
Teacher evaluation of students: two measures were administered to assess (a) teachers' expectation for students (before administration of tests), (b) teachers' rating of academic ability on a five-point Likert-type response scale, and (c) teachers' rating of students' academic ability a second time, four weeks later.

Differential Aptitude Test (DAT): The Verbal Reasoning and Numerical Ability scales were administered to the students.

Student self-evaluations: Two or three weeks after the DAT tests, the students were asked to respond to 'How well did you do on the DAT?'. This was repeated one to two weeks later as a retest.

Student grades in English and mathematics were obtained from the school records.

Data analysis
Stability coefficients (test-retest estimates) were calculated for the students, self-evaluations. Accuracy of student self-evaluations were estimated by comparing actual and self-evaluations on the DAT scores.

The relationship between student self-evaluations, teacher assessment and course performance variables and actual performance on the DAT, correlation and multiple regression analyses were used.

Authors' findings
Reliability of student self-assessment
Test-retest reliability coefficients for verbal ability were strong for grades 9 to 12 (0.82 to 0.96) but less so for grade 8 (0.56)
Self-estimates were less reliable for mathematics, although still high for the twelfth-grade students in mathematics.

Relationship of self-assessment to actual performance:
Correlations were 'modest to moderate' although reaching significance at the 0.05 level in 11 out of 12 cases. The proportion of 'direct hits' ranged from 0.25 for 10th grade numerical ability to 0.74 for 11th grade numerical ability.

Overall, the strongest correlates with the criterion variables (Verbal reasoning, VR, and numerical ability, NA) were made by the teachers' evaluations (either the probability of success or the academic ability) and by students' self-estimates.

Teacher assessments estimates correlated with actual performance fairly stably across grades. The teachers' estimates were moderate predictors for both verbal and numerical ability (coefficients 0.59 for VR and 0.47 for NA).

Eleventh-grade multiple regression results (reported only for the 11th grade, where there were sufficient numbers) showed that three variables were effective predictors of VR scores: teacher assessments of academic ability, student self-evaluation and grade averages. All were significant predictors, with the self-assessment just slightly stronger.

Authors' conclusions
The authors advise caution in interpreting the findings. They conclude: 'based on the regression analyses for the eleventh-grade sample, that student self-assessments and teacher perceptions of student ability and probability of success are important moderately valid predictors of academic performance - at least in an academic setting in which students are at risk of dropping out of school' (p 268).

For all grades, teacher assessments achieved moderate to strong correlations with student performance on the verbal ability test and slightly less strength for the numerical test. These findings suggest that teachers may relay upon a perceived 'verbal competency' dimension in judging a student's academic ability as well as for estimating a student's potential for success in completing a given course of study.

'In closing, for a sample of Appalachian 11th graders, students self-estimates of verbal and mathematical ability appear to contribute unique information as an indicator of actual performance when considered jointly with teacher assessments and grade performance average' (p 269).