REVIEW

June 2002

EPPI-Centre

A systematic review of the impact of summative assessment and tests on students' motivation for learning

Review conducted by the Assessment and Learning Research Synthesis Group

Evidence for Policy and Practice
Information and Co-ordinating Centre

The EPPI-Centre is part of the Social Science Research Unit, Institute of Education, University of London
http://eppi.ioe.ac.uk/

AUTHORS AND INSTITUTIONAL BASES

This work is a review of the Assessment and Learning Research Synthesis Group (ALRSG). It was conducted following the procedures for systematic review developed by the EPPI-Centre and in collaboration with David Gough and Dina Kiwan, and with help from other members of the EPPI-Centre education team.

Principal authors

Wynne Harlen, Graduate School of Education, University of Bristol.
Ruth Deakin Crick, Graduate School of Education, University of Bristol.

ALRSG members

Professor Patricia Broadfoot, University of Bristol
Professor Richard Daugherty, University of Wales, Aberystwyth
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
(The above are also members of the Assessment Review Group.)

Mr P Dudley, Head of School Improvement and Lifelong Learning, Redbridge, and member of AAIA
Mr R Bevan, Deputy Headteacher, King Edward VI Grammar School, Chelmsford
Ms P Rayner, Headteacher, Caldecote Primary School, Cambridge
Dr Ruth Deakin Crick, Researcher

EPPI-Centre members

Dr David Gough, Deputy Director
Ms Dina Kiwan, Education Research Officer

Expert advisers to the ALRSG

The ALRSG is advised by the following international experts:

Dr Steven Bakker, ETS International, The Netherlands
Dr Dennis Bartels, 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

Thanks

The authors would like to thank the above and members of the EPPI-Centre team for their guidance and support during the conduct of the review.

Acknowledgements

This review was carried out with funding from the Nuffield Foundation and from the EPPI-Centre (Evidence for Policy and Practice Information and Coordinating Centre).

Submission date: 30/05/2002

This report should be cited as: Harlen W, Deakin Crick R (2002).
A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). In: *Research Evidence in Education Library*. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

* This review has been updated since it was published originally. This update involves minor amendments only and has not changed the substantive findings of the review.

TABLE OF CONTENTS

SUMMARY

Background

The current widespread use of summative assessment and tests is supported by a range of arguments. The points made include that not only do tests indicate standards to be aimed for and enable these standards to be monitored, but that they also raise standards. Proponents claim that tests cause students, as well as teachers and schools, to put more effort into their work on account of the rewards and penalties that can be applied on the basis of the results of tests. In opposition to these arguments is the claim that increase in scores is mainly the consequence of familiarization with the tests and of teaching directed specifically towards answering the questions, rather than developing the skills and knowledge intended in the curriculum. It is argued that tests motivate only some students and increase the gap between higher and lower achieving students; moreover, tests motivate even the highest achieving students towards performance goals rather than to learning goals, as required for continuing learning.

This systematic review was prompted by concern to identify the impact of summative assessment and testing, which has burgeoned in many countries in the past decade, on students' motivation for learning. Whilst the impact of testing on teachers, teaching and students' achievement has been well researched and represented in reviews of research, much less attention has been given to its impact on the affective and conative (mental activity) outcomes of education. The current widely embraced aim of developing in today's students the capacity to continue learning beyond the years of schooling into lifelong learning means that, if some assessment practices are reducing motivation for learning, there is clearly a cause for concern. The purpose of the review was therefore to identify and synthesise research evidence about the impact of summative assessment on motivation for learning.

Definition of terms

Assessment is a term that covers any activity in which evidence of learning is collected in a planned and systematic way, and is used to make a judgment about learning. If the purpose is to help in decisions about how to advance learning and the judgement is about the next steps in learning and how to take them, then the assessment is formative in function. If the purpose is to summarise the learning that had taken place in order to grade, certificate or record progress, then the assessment is summative in function. When summative assessment is used for making decisions that affect the status or future of students, teachers or schools (that is, 'high stakes'), the demand for reliability of measures often means that tests are used in order closely to control the nature of the information and the conditions in which it is collected.

Motivation is a complex concept concerned with the drive, incentive or energy to do something. Motivation is not a single entity but embraces, for example, effort, self-efficacy, self-regulation, interest, locus of control, self-

esteem, goal orientation and learning disposition. Learning, too, is a complex phenomenon that cannot be conceived as a single entity but is best understood as a field or as an ecological composite. The American Psychological Association's 'Learner Centered Psychological Principles' include 14 factors that influence learning and learners. These include cognitive and metacognitive factors, motivational and affective factors, developmental and social factors, and individual difference factors. Thus for the purpose of this review, motivation for learning is understood to be a form of energy which is experienced by learners and which drives their capacity to learn, adapt and change in response to internal and external stimuli. It is closely identified with the 'will to learn', which determines the effort that a learner will put into a task.

There are different ways in which the energy, or the will, to learn can be motivated and it is particularly important to distinguish between intrinsic and extrinsic motivation. Those who learn in order to gain an extrinsic reward are unlikely to continue learning once the reward is obtained or the penalty avoided, and they will give up earlier if reward seems unobtainable. For continued learning, the motive needs to be intrinsic, the reward being in the process of learning and in the recognition of being in control of, and responsible, for one's own learning.

Aims of the review and review questions

Aims

The aims of the review were as follows:

1. To conduct a systematic review of research evidence of the impact of summative assessment and testing on students' motivation for learning
2. To determine the conditions and processes (including teaching) associated with summative assessment and testing having a positive or a negative impact on students' motivation for learning
3. To identify actions that could be taken to increase the positive and decrease the negative impact of assessment on students' motivation for learning
4. To consider evidence relating motivation for learning to learning achievements and learning strategies
5. To make recommendations for policy and practice based on these findings
6. To identify questions that need to be addressed by research so that decisions on policy and practice in summative assessment can be evidence-based

Review questions

Thus the review was designed to identify and synthesise research relevant to the question:

- What is the evidence of the impact of summative assessment and testing on students' motivation for learning?

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

- How does any impact vary with the characteristics of the students and the conditions of the assessment or testing?
- In those studies where impact on students has been reported, what is the evidence of impact on teachers and teaching?
- What actions in what circumstances would increase the positive and decrease the negative impact on students of summative testing and assessment programmes? In particular, what is the evidence that any impact is increased by 'raising' the stakes?
- What are the implications for assessment policy and practice of these findings?

Methods

The review was conducted using the procedures for systematic review of research in education being developed by the EPPI-Centre. A wide-ranging search was carried out for studies, written in English, of assessment for summative purposes in schools for students between the ages of 4 and 19, and which reported on aspects of students' motivation for learning. The search for studies involved searching relevant electronic databases and journals online, following up citations in other reviews, handsearching journals held in the library, and using personal contacts. Inclusion and exclusion criteria were applied to abstracts before full texts were read and labelled, using a core set of keywords and additional keywords specific to the review. This process resulted in some further studies being excluded. The remaining studies were analysed in depth using the *Guidelines for Extracting Data and Assessing quality of Primary Studies in Educational Research,* Version 0.94 (EPPI-Centre, 2001). Judgements were made as to the weight of evidence relevant to the review provided by each study.

Lengthy consideration was given to ways in which the findings of different studies could be brought together to form conclusions. None of the studies dealt with all the variables included in the concept of motivation for learning but they could be grouped according to the particular outcomes that were investigated. These outcomes fell into three distinct groups, central to motivation for learning. Expressed from a learner's perspective, these are as follows:

1. What I feel and think about myself as a learner
2. The energy I have for the task
3. How I perceive my capacity to undertake the task

The findings relating to the main review question are reported under these headings. Judgements were made about each study in relation to methodological soundness, appropriateness of the study type and relevance to the focus of the review. In the synthesis, greater weight was accorded to those studies rated most highly on these counts.

Results

The initial search resulted in the identification of 183 potentially relevant studies. The successive stage in the systematic review process involved excluding some studies at various stages, for reasons that were documented. It resulted in 19 studies being identified as directly addressing the review question; these included 13 outcome evaluations (three

randomised controlled trials, three case control designs, three post-test and four of other designs), three descriptive studies and three process evaluations.

Evidence of impact

Between them, the identified studies considered a number of the component aspects of motivation, but none considered all. The following main findings emerged from studies providing high-weight evidence:

- After the introduction of the National Curriculum Tests in England, low-achieving pupils had lower self-esteem than higher-achieving pupils, whilst beforehand there was no correlation between self-esteem and achievement.
- When passing tests is high stakes, teachers adopt a teaching style which emphasises transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences.
- Repeated practice tests reinforce the low self-image of the lower-achieving students.
- Tests can influence teachers' classroom assessment which may be interpreted by students as purely summative, regardless of the teacher's intentions, possibly as a result of teachers' over-concern with performance rather than process.
- Students are aware of a performance ethos in the classroom and that the tests give only a narrow view of what they can do.
- Students dislike high-stakes tests, show high levels of test anxiety (particularly girls) and prefer other forms of assessment.
- Teachers have a key role in supporting students to put effort into their learning activities.
- Feedback on assessments has an important role in determining further learning. Students are influenced by feedback from earlier performance on similar tasks in relation to the effort they invest in further tasks.
- Teacher feedback that is ego-involving rather than task-involving can influence the effort students put into further learning and their orientation towards performance rather than learning goals.
- High-stakes assessment can create a classroom climate in which transmission teaching and highly structured activities predominate and which favour only those students with certain learning dispositions.
- High-stakes tests can become the rationale for all that is done in classrooms, permeating teacher-initiated assessment interactions.
- Goal orientations are linked to effort and self-efficacy.
- Teacher collegiality is important in creating an assessment ethos that supports students' feelings of self-efficacy and effort.
- An education system that puts great emphasis on evaluation produces students with strong extrinsic orientation towards grades and social status.

Evidence from studies providing medium-weight evidence:

- The state-mandated tests in the US lower self-esteem for 'at risk' students.

- Low-achieving children can improve their achievement with the help of supportive teachers or other adults.
- Interest and effort are encouraged in classrooms which encourage self-regulated learning by providing students with an element of choice, control over challenge and opportunities to work collaboratively.
- Teachers can influence the criteria that students use in self-assessment of their work.

Differences relating to age, gender and level of achievement

Older students (that is, aged 11 and above) are more likely than younger ones to have a better understanding of simple grades are less likely to report teachers' grades as being fair but attached more importance to them. Older students are more likely to attribute relative success to effort and ability, whilst younger ones attribute it to external factors or practice. Older students are also more likely to focus on performance outcomes (that is, scores and levels) rather than learning processes.

Lower-achieving older students were more likely than younger ones to minimise effort and respond to tests randomly or by guessing. There was no evidence of age differences in test-taking strategies (checking, monitoring time, etc.). Instead of motivation and test familiarity increasing with age, older students feel more resentment, anxiety, cynicism and mistrust of standardised achievement tests.

Lower-achieving students are doubly disadvantaged by summative assessment. Being labelled as failures has an impact not just on current feelings about their ability to learn, but lowers further their already low self-esteem and reduces the chance of future effort and success. Only when low achievers have a high level of support (from school or home), which shows them how to improve, do some escape from this vicious circle.

Results of tests which are 'high stakes' for individual students, such as the 11+ in Northern Ireland, have been found to have a particularly strong and devastating impact on those who receive low grades. However, tests which are high stakes more for schools than for students (such as the National Curriculum tests in England and state-mandated tests in the US) hardly have less impact. Students are aware of repeated practice tests and the narrowing of the curriculum and only those confident of success enjoy the tests. In taking tests, high achievers are more persistent, use appropriate test taking strategies and have more positive self-perceptions than low achievers. Low achievers become overwhelmed by assessments and demotivated by constant evidence of their low achievement, thus further increasing the gap between low- and high-achieving students. A greater emphasis on summative assessment thus brings about increased differentiation.

Girls were reported as expressing more test anxiety than boys. Girls also make more internal attributions of success or failure than boys, with consequences for their self-esteem.

The effect of the conditions of testing

The conditions that affect the impact of summative assessment relate to the degree of self-efficacy of students, the extent to which their effort is intrinsically or extrinsically motivated, the encouragement of self-regulation and self-evaluation, and the pressure imposed by adults outside the school. Feedback has a central role since self-efficacy is judged from performance in previous tasks of the same kind. If students have experienced success in earlier performance, they are more likely to feel able to succeed in a new task. Feedback that focuses on the task is associated with greater interest and effort, whereas feedback that is ego-involving rather than task-involving is associated with an orientation to performance goals.

Teachers' own class-testing practices can help to increase self-efficacy, if teachers explain the purpose and expectations of their tests and provide feedback. Further, a school's 'assessment culture' influences students' feelings of self-efficacy and effort. Collegiality – meaning constructive discussion of testing and the development of desirable assessment practice in the school – has a positive effect, whilst a focus on performance outcomes has a negative effect.

The degree to which learners are able to regulate their own learning also appears to favour students' interest and to promote focus on the intrinsic features of their work. Students who have some control over their work by being given choice and encouragement to evaluate their own work value the significant content features of their work, rather than whether it was correct or not. Thus classrooms that allow more self-regulation promote change in the criteria students use in self-evaluation.

When test scores are a source of pride to parents and the community, pressure is brought to bear on the school for high scores. Similarly, parents bring pressure on their children when the result has consequences for attendance at high social status schools. For many students, this increases their anxiety, even though they recognise their parents as being supportive.

The effect on teachers and teaching

High-weight evidence from studies reporting on the effect of tests on teachers and teaching in addition to impact on students' motivation indicates that when passing tests is high stakes, teachers adopt a teaching style which emphasises transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences. External tests have a constricting effect on the curriculum, resulting in an emphasis on subjects tested at the expense of creativity and personal and social development. High-stakes tests often result in a great deal of time being spent on practice tests, the valuing of test performance and undervaluing of other student achievements, with teachers' own assessment becoming summative in function rather than formative.

Increasing the positive and decreasing the negative impact

Although the study findings pointed to negative impacts of summative assessment on aspects of motivation for learning, they also indicated ways in which these could be ameliorated so that learners as well as teachers can benefit from summative assessment. High-weight evidence suggests that practice in summative assessment could be improved by the following:

- Promoting learning goal orientation rather than performance orientation
- Cultivating intrinsic interest in the subject and putting less emphasis on grades
- Teaching approaches that encourage self-regulated learning (including collaboration among students) and cater for a range of learning styles
- Providing explanations of the purpose of assessment and providing feedback that can help further learning
- Establishing a school climate of constructive discourse about assessment among teachers, and between teachers and students
- Developing a constructive and supportive school ethos in relation to tests
- Ensuring that the demands of the tests are consistent with the expectations of teachers and the capabilities of the students
- Involving students in decisions about testing
- Developing students' self-assessment skills and use of learning rather than performance criteria as part of a classroom environment that promotes self-regulated learning
- Using assessment to convey a sense of learning progress to students

Implications for assessment practice and policy

In order to explore the implications of the review as fully as possible, the review methodology included a consultation conference with invited policy-makers and practitioners. Some of the messages below are derived directly from the research studies whilst others emerged from discussion of the review findings and reference to current practice in the UK at the consultation conference.

Practice

- Reduce the narrowing impact on the curriculum and on teaching methods by professional development that emphasises learning goals and learner-centred teaching approaches.
- Share and emphasise learning goals, rather than performance goals, with students and provide feedback to students in relation to these goals.
- Share in developing and implementing a school-wide policy that includes assessment both *for* learning (formative) and *of* learning (summative), and ensure that the purpose of all assessment is clear to all involved, including parents and students.
- Develop students' understanding of the goals of their learning, the criteria by which they are assessed and their ability to assess their own work
- Implement strategies for encouraging self-regulation in learning and positive inter-personal relationships.

- Avoid comparisons between students based on test results.
- Present assessment realistically, as a process which is inherently imprecise and reflexive, with results that have to be regarded as tentative and indicative rather than definitive.

Policy

- Recognise that current high-stakes testing is providing information about students' attainment by reducing motivation that is of questionable validity.
- Recognise the importance of the various components of motivation for students' attainments in education. Empirical evidence shows that these are positively related to attainment. For example, the OECD/PISA (2001) provides firm evidence that achievement of literacy is positively related to students' interest in their learning, the extent to which their learning strategies help them to develop understanding through linking to existing knowledge instead of just memorising, and the extent to which they feel in control of their learning.
- Provide professional development, particularly for senior school management, aimed at enabling schools to develop a range of assessment strategies and using summative information of different kinds for improving the learning of their students. Current training focuses too narrowly on the use of test scores, accountability and target-setting; it needs to be more learner-focused.
- For summative purposes in reporting on individual students, move towards testing students when their teachers judge them to be ready to show their achievement at a certain level, thus minimising experience of failure and its impact on self-esteem.
- Ensure that the criteria used in school evaluation (including self-evaluation) make explicit reference to a full range of subjects; include moral, spiritual and cultural as well as cognitive aims; and range across an appropriate variety of teaching methods and learning outcomes.
- Develop schools' self-evaluation practices, including teachers' assessment skills, through targeted professional development.
- For tracking national standards, sample students rather than test all and use a wider range of test forms and items.
- Quantify the 'cost' of current practice, including teaching time taken up with testing and practice testing; the additional workloads to teachers' of extra marking; in addition to the cost of the tests and their development.
- Use test development expertise to create new tests and assessment that will enable all valued outcomes of education, including creativity and learning to learn to be assessed.
- Reduce the 'stakes' of summative assessment by avoiding comparisons among schools in terms of test results and end the practice of basing targets only on test results.

Other outcomes of the review were the identification of further research required in this area particularly to extend the research base in relation to outcomes of education that are particularly important for lifelong learning, and a clarification and development of the methodology of systematic reviewing applied to educational research.

1. BACKGROUND

1.1 The increase in summative assessment and tests

Many developments in education policy are designed to raise standards of students' achievement; assessment policy is an important element. There are two competing claims for ways in which assessment can raise standards. On the one hand, there is a common sense assumption, widespread among both the educational community and parents, that summative assessment in the form of tests and examinations, is a key source of motivation for learning. In England, as in many states of the USA, where assessment for summative purposes has burgeoned in the past decade, an increase in test scores year on year has been found and this has been attributed, as least in part, to the implementation of tests.

The growth of external tests in the US has been charted by Clarke, Madaus, Horn and Ramos (2000). They report that the number of states using standards-based tests rose to 47 in 1998, an increase of 40% in just three years. They also comment on the trends in the form of tests, a feature which cannot be neglected when considering the impact on teaching and learning. They report that, although multiple choice tests, which dominated in the first half of the 20th century, were challenged by the move towards 'alternative assessment' in the 1980s, the drive for efficiency in testing and the use of machine marking means that a large number of states continue to use, or have reverted to the use of, multiple choice tests, although some supplement these with short-answer questions. The authors also point out the very large commercial industry that has grown up around testing, with inevitable pressures to maintain a high level of use of tests. Schools use a variety of standardised tests in addition to the mandated standards-based.

In England, too, there is test-inflation. A survey by the Qualifications and Curriculum Authority conducted in 2000 found that the introduction of National Curriclum Tests brought with it an increase, not a decrease, in use of other tests. It is estimated that the average students in England takes 105 tests between the ages of 4 and 18 (Professional Association of Teachers, 2002).

Thus summative assessment has become for most students in many Western countries, not a once-a-year event which in comparison with daily interactions with teachers might be considered to have a minor role in determining their 'faith in themselves as learners' (Stiggins, 2001 p46), but rather a frequent experience which may have an undesirable effect on motivation for learning. Moreover, research shows that this effect is greater for the less successful pupils and thus tends to widen the gap between higher and lower achieving pupils (Madaus, 1992).

1.2 Assessment and raising standards

A strong case can be made for the motivational role of external examinations and tests. Kellaghan, Madaus and Raczek (1996) identified six propositions put forward in favour of this role. These are: that tests and examinations

indicate standards; that high ('world class') standards can be demanded; that they exemplify to students what they have to learn; that rewards and penalties can be applied to the results; that students will put effort into school work in order to pass tests; that this will be the case for all students. Most, if not all, of these propositions underpin summative assessment programmes such as state-mandated tests in the US, the national examination systems for 16 to 19 year olds in the UK and in many other countries, and the national curriculum tests in England and Wales.

On the other hand, there is evidence from research into classroom assessment, cogently revealed in the review by Black and Wiliam (1998), that formative assessment – assessment that is integral to teaching and designed to help learning – raises standards. Using a well documented, although not strictly systematic, methodology, Black and Wiliam surveyed the evidence about classroom assessment by teachers and its impact on learning. From an initial 681 research studies they summarised findings from about 250. Their conclusions refer to the use of assessment for formative purposes, which they find to have a wholly positive impact on learning when it includes certain key features which emerged from the studies. These features include particular forms of feedback, the involvement of students in self-assessment and the use of assessment in modifying teaching. Black and Wiliam acknowledged that such practices require large shifts in teachers' perceptions of their roles in relation to their students, but that considerable gains in achievements are possible as a result.

On the face of it, if the claims for both summative assessment and formative assessment are valid, the two could co-exist in educational practice, combining to raise standards for all students. A widely expressed view of educators who have conducted research into summative assessment is that the increase of test scores over time is due to greater familiarity of teachers and pupils with the tests rather than increasing learning (e.g. Koretz 1988, 1991; Linn, 2000; Kohn, 2000). Further, the use of test scores and examinations for purposes which affect the status or future of students, teachers or schools (that is, are 'high stakes') results in teachers focusing teaching on the test content, training students in how to pass tests, and adopting teaching styles which do not match the preferred learning style of many students (Johnston and McClune, 2000). In these circumstances, teachers make little use of assessment formatively, to help the learning process (Pollard *et al.*, 2000; Osborn, 2000; Broadfoot *et al.*, 1998). In other words, high-stakes summative assessment squeezes out formative assessment. Black and Wiliam also warned that 'the context of national and local requirements for certification and accountability will exert a powerful influence on (formative assessment) practice' (Black and Wiliam, 1998, p. 20).

1.3 Impact on students and teachers

The impact of summative assessment on teachers and teaching has been well researched and represented in the reviews of Crooks (1988), Linn (1982) and Shepard (1991). Crooks looked at the impact of assessment on students, including self-efficacy, intrinsic motivation and attribution of success or failure. He found evidence of the importance of motivational aspect in relation to classroom assessment; that the use of extrinsic motivation is problematic and that intrinsic motivation and self-regulated learning is important to continued learning both within and without school. He reviewed the potentially positive

role of classroom assessment, for example, in helping students to focus their learning, but also concluded that test anxiety has a debilitating effect on achievement and that this could be reduced by avoiding comparisons between students and the use of letter grades.

Gordon and Reese (1997) reported evidence that teachers can train students to pass any kind of test, even those intending to assess higher thinking skills, frustrating those who consider that teaching to well designed tests can influence teaching in positive directions (e.g.Yeh, 2001). Kellaghan *et al.* (1996) expressed doubts that the aims of the education reform which emphasises higher level thinking and problem-solving skills are compatible with the programmes of high stakes testing. They traced the mechanism for orienting students towards performance goals to the way in which students are prepared for high stakes tests. The research they reviewed also undermined the claim that better tests will lead to better teaching and learning.

> 'Proponents of a system of high-stakes examinations will argue that if we get the right kinds of tests – ones worth teaching to and preparing for – then test-preparation practices will lead to the development of the valued skills purportedly measured by the test. However, we believe that this argument seriously underestimates the ability of test preparation to corrupt the very construct the test is trying to measure. …An important implication of this is that when such corruption occurs, inferences from the test to the original domain of interest – which if the educational reform language is to be believed is the domain of higher-order thinking skills and habits of learning – will no longer be valid.'
> (Kellaghan *et al.*, 1996, p. 53)

The impact of summative assessment on students' motivation for learning can be both direct and indirect. A direct impact can be through inducing test anxiety and the effect of low scores on self-esteem and perceptions of themselves as learners; an indirect impact can be through the effect on their teachers and the curriculum. Any negative impact on motivation for learning is clearly highly undesirable, particularly at a time when the importance of learning to learn and lifelong learning is widely embraced. Thus it has been argued that a rise in test scores may be accompanied by unintended negative outcomes which have serious consequences for current generations of students.

This review was undertaken in order to identify research evidence on this matter and to report on what is known about the impact of summative assessment on motivation. First, we have to be clear about what is included in the concept of motivation for learning.

1.4 Motivation for learning

Motivation for learning is a complex overarching concept, which constitutes a range of psychosocial factors both internal to the learner and present in the learner's social and natural environment. Psychologists apply the term 'motivation' to 'the conditions and processes that account for the arousal, direction, magnitude and maintenance of effort' (Katzell and Thompson, 1990, p. 144) which indicates this complexity. Johnston (1996) argues that the 'will to learn' is at the very heart of the learning process and that this is very closely aligned with the concept of motivation. She argues that the will to learn is

derived from a person's sense of deep meaning, or sense of purpose, and can be described as the energy to act on what is meaningful. The will to learn is related to the degree to which the learner is prepared to invest effort in the learning process, and is that which engages their motivation to process, perform and develop as a learner over time.

Common to many theories which have been built around the concept of motivation is reference to goal orientation. People who commit themselves to a goal will direct their attention towards actions that help them to attain that goal and away from other actions. Research indicates that students with learning goals show more evidence of superior learning strategies, have a higher sense of competence as learners, show greater interest in school work and have more positive attitudes to school than do students with performance or achievement goals (Ames, 1990; Dweck, 1992).

There are many reasons why a goal may or may not be embraced. In their review of research evidence Kellaghan *et al.* (1996) suggest that these include, firstly, the need for an individual to comprehend the goal; secondly, that the goal needs to be reachable yet challenging; thirdly, that individuals should believe that their efforts to reach the goal will be successful; and, fourthly, that attainment of the goal should lead to actual benefit for the individual.

Attribution theory, concerning the reasons one gives for personal characteristics or behaviour, contributes to our understanding of motivation for learning, because a person's perceptions of the causes of success and failure are of central importance in the development of effective learning. Causes have three dimensions. The first is *locus*, whether causes are perceived to originate from within the person or externally. The second is *stability,* whether the causes are perceived to be constant or to vary over time. The third has to do with *controllability* whether the individual perceives that she or he can influence the causes of success or failure.

Ability and effort are two frequently used causes of success or failure at a learning task. Both are internal to the learner, but perceptions of their stability and controllability vary among learners and teachers. Learners who attribute success to ability, which they perceive as stable and uncontrollable, are likely to respond positively to summative assessments, whereas learners who attribute failure to ability, which they perceive as stable and uncontrollable, are likely to respond negatively to summative assessment. Concomitantly, learners who attribute success to effort, and who perceive ability to be changeable and controllable are likely to deal with failure constructively, and to persevere with the learning task (Schunk, 1991). All of these factors contribute to a learners' sense of efficacy in learning: their capacity to learn and to go on learning.

1.4.1 Intrinsic and extrinsic motivation

Many educational psychologists and researchers distinguish between intrinsic and extrinsic motivation. Intrinsic motivation means that learners find interest and satisfaction in what they learn and in the learning process itself; it leads to self-motivated and continued learning. Learners who are 'motivated from within' recognise their own role in learning and so take responsibility for it. These are the learners who will seek out information, identify their own learning goals and persevere, knowing that what they achieve depends on

their own effort. Such motivation is clearly needed for lifelong learning. Extrinsic motivation describes the behaviour of learners who engage in learning because it is a means to an end that has little to do with the content of what is learned. The incentive for learning is found in rewards such as certification, merit marks, prizes or in avoiding the consequences of failure. Not only does this mean that learning may stop, or at least that effort is decreased, in the absence of such external incentives, it also means that what is learned is closely targeted at behaviour which is rewarded.

Some researchers have gone beyond the identification of intrinsic and extrinsic motivation. For example, McMeniman (1989) identifies 'achievement' motivation, which is concerned not with external rewards, nor with the pleasure of learning, but with the desire to achieve success in comparison with others. Research indicates that competition with others and norm-referenced assessment works best for successful students (Natriello, 1987).There are also various types of extrinsic motivation identified by Rigby *et al.* (1992) which are related to a dimension of different degrees of self-determination.

There is a considerable body of opinion and evidence that suggests that different kinds of motivation are associated with different learning strategies. For example, intrinsic motivation is associated with levels of engagement that lead to development of conceptual understanding and higher level thinking skills (Kellaghan *et al.*, 1996). The review by Crooks (1988) also drew attention to research that indicated the value of intrinsic motivation and the problems associated with extrinsic motivation in tending to lead to 'shallow' rather than 'deep' learning.

Rigby *et al.* (1992) argue that a simple dichotomy between intrinsic and extrinsic motivation is misleading when trying to explain learning because individuals will seek to internalise and integrate their extrinsically motivated behaviours. They suggest instead that there is a continuum of self-determination, which begins with *external regulation*. The second point on the continuum is *introjected regulation*, which involves motivation by internal pressures such as self-esteem. The third point is *identified regulation*, where the individual has come to value the behaviour and has accepted the regulatory processes. The final and most autonomous form of extrinsic motivation is called *integrated regulation* when the regulatory process is fully integrated in the individual's sense of self, identity and value system.

1.4.2 The role of rewards

A good deal of attention had been given to the effect of rewards on motivation. Kohn (1993), for example has conducted experimental studies which he interprets as showing that associating a particular behaviour with a reward decreases the likelihood of the behaviour being continued voluntarily if not again rewarded. Others have concluded from similar experimental studies that attention is narrowly focused on which is required to obtain the reward. However opinions differ as to the dependability of the research and Kellaghan *et al.* (1996) commented that the results of experimental studies are not clear-cut and findings vary considerably with circumstances.

The meta-analysis in a systematic review by Deci, Koestner and Ryan (1999) of 128 studies of the effects of extrinsic rewards on intrinsic motivation appear to show clearly that such rewards undermined intrinsic motivation across a wide range of activities, populations and types of reward. However, the

conclusions of Deci *et al*. have been challenged by Hidi (2000) who pointed out that these were drawn from studies that only related to activities that were interesting (as defined by the researcher), and excluded uninteresting tasks. There was also relatively short-term engagement of learners with the activities. In a separate analysis of a smaller number of uninteresting tasks, rewards did not reduce intrinsic motivation. Hidi and Harackiewicz (2000) argue that 'the effects of external rewards may depend on the complexity of the activity and the length of involvement. More specifically, a combination of intrinsic rewards inherent in interesting activities and external rewards, particularly those that provide performance feedback, may be required to maintain individuals' engagement across complex and often difficult – perhaps painful – periods of learning' (ibid p. 159). From their own review of research on the role of interests and goals on achievement, Hidi and Harackiewicz, concluded that the dichotomy between extrinsic and extrinsic motivation is unhelpful and that it is time to seek 'optimal combinations'. This may be particularly necessary for students lacking interest and intrinsic motivation for academic studies.

Ames (1992) review of the research on the impact of rewards in the context of summative classroom assessment supported the findings of Deci and Ryan, (1985), but added that other research shows that rewards can enhance achievement goals when it is contingent on effort, on progress in relation to short-term goals or on 'meaningful aspects of performance'.

1.5 Relationship between assessment and aspects of motivation

From the evidence cited above, it appears that both the extent of learning and the quality of that learning may depend on the type of motivation. The role that assessment plays in promoting intrinsic or extrinsic motivation is a central one and hinges on the concept of the perceived locus of control. Learning has to be done by the learners, for no-one else can do it for them; therefore, the more the learner is in control, the more likely that there is effort put into learning. The review of Deci and Ryan (1985) provides research evidence that assessment of the kind that takes away control from the learners reduces intrinsic motivation and leads to 'surface' learning.

Natriello (1987) emphasised the importance of distinguishing between the different purposes of assessment and concluded that 'an evaluation system for the purpose of enhancing student motivation might involve a differentiated task structure in the classroom, a mix of more and less predictable tasks, clearly articulated criteria, challenging yet attainable, self-referenced standards, relatively frequent collection of information, appraisals that truly reflect student effort and performance and differentiated and encouraging feedback. An evaluation system designed for purposes of certification would look quite different' (p. 171). Whilst reviewing research which indicates that clarifying assessment criteria for students is associated with high levels of skill, self-efficacy and rapid problem-solving, he pointed out that too much emphasis on explicit criteria can have the undesirable effect of encouraging students to limit attention to what is being assessed.

Ames' review of 1992 was concerned with looking at achievement goals and identifying the situations and instructional strategies that lead to motivation towards desired goals. She contrasted learning goals (also known as task-involved or mastery goals) with performance goals (also known as ego-

involved goals). In searching for conditions which affect students' motivation for learning, she cited research which indicates that social comparisons have a strong role in this respect. Students who are compared unfavourably and publicly with their peers hold low self-esteem in relation to learning, avoid risks and use less effective and more superficial learning strategies. Not only do their own perceptions of themselves as learners suffer but this perception becomes shared by their peers. She cites Grolnick and Ryan's (1987) findings that when assessment is perceived as 'an attempt to control rather than inform, meta-cognitive processes are short-circuited' (p. 265).

A review by McDonald (2001) was specifically focused on test anxiety and its impact on students' performance. His concern was to look at evidence relating to students at school, since he notes that conflicting conclusions about the impact of test anxiety on performance may have resulted from many studies having been carried out in experimental situations with those who have left compulsory education. He found studies difficult to synthesise on account of the different instruments used to assess test anxiety. Where there was a distinction between general fears and test anxiety (fear of negative assessment), it was found that, whilst the former decrease with age, the latter increase with age. Females were found to score more highly on test anxiety than males. In relation to performance, there was considerable evidence from a range of countries and across academic subjects of a negative relation between test anxiety and test performance. Although there were also studies which reported no relationship, McDonald concludes that overall the influence is negative and large enough to make the difference between passing and failing a test for at least one fifth of the students. Thus this review supports the importance of paying attention to test anxiety.

1.6 Differential impact on minority students

Two reviews – by Madaus and Clarke (1998) and McNeil and Valenzuela (2000) – were presented at a conference on High Stakes Testing K-12 held at Harvard University in December 1998. They had specific focus on research relating to issues of high stakes testing in the US. Madaus and Clarke focused on the impact of high-stakes testing on minority students, drawing mainly on research conducted at Boston College's Center for the Study of Testing, Evaluation and Educational Policy. They use the research not only to identify the existence of impact but also how high-stakes testing comes to influence what is taught and learned. They point out that such influence is deliberate in a context of 'measurement-driven instruction' and show that teachers use past examination papers to define the curriculum, paying attention not just to the content but also the form of the test. They discuss the impact on student motivation and on student dropout rate. They draw the following conclusions:

- High-stakes, high-standards tests do not have a markedly positive effect on teaching and learning in the classroom.
- High-stakes tests do not motivate the unmotivated.
- Contrary to popular belief, 'authentic' forms of high-stakes assessment are not a more equitable way to assess the progress of students who differ in race, culture, native language or gender.
- High-stakes testing programs have been shown to increase high school dropout rates – particularly among minority student populations. (Madaus and Clarke, 1998, p. 1)

McNeil and Valenzuela (2000) reviewed evidence of the impact of high-stakes testing in general and of the Texas Assessment of Academic Skills (TAAS) in particular. Like Madaus and Clarke, their focus was on the impact on minority and economically disadvantaged students. They present an analysis of studies from which they conclude that

> 'behind the rhetoric of rising test scores are a growing set of classroom practices in which test-prep activities are usurping a substantive curriculum. These practices are more widespread in those school where administrator pay is tied to test scores and where test scores have been historically low'
> (McNeil and Valenzuela, 2000, p. 2).

In such schools, mostly attended by African-American and Latino students, the pressure has meant that 'a regular education has been supplanted by activities whose sole purpose is to raise test scores on this particular test.' (p. 2). McNeil and Valenzuela highlight the distortion of educational expenditure: away from high quality curriculum resources towards test-preparation materials which have little educational benefit beyond the test.

2. AIMS OF THE REVIEW AND REVIEW QUESTION

2.1 Aims of the review

The purpose of this review was to address, through a systematic synthesis of the evidence from research, the important relationship between students' motivation for learning and summative testing and assessment. The overall aim was to identify the research evidence of the impact of summative assessment on the various component variables that together constitute 'motivation', or the energy to learn. In addition, in order to make recommendations for policy and practice, it is important to understand how any impact is brought about. Thus the review sought to identify the conditions and processes involved in creating an impact, including the role of teachers and teaching.

The aims of the review described in this report were:

- To conduct a systematic review of research evidence of the impact of summative assessment and testing on students' motivation for learning
- To determine the conditions and processes (including teaching) associated with summative assessment and testing having a positive or a negative impact on students' motivation for learning
- To identify actions that could be taken to increase the positive and decrease the negative impact of assessment on students' motivation for learning
- To consider evidence relating motivation for learning to learning achievements and learning strategies
- To make recommendations for policy and practice based on these findings.
- To identify questions that need to be addressed by research so that decisions on policy and practice in summative assessment can be evidence-based

The review includes process and outcome studies of interventions relating to various forms of summative assessment, including standardised tests, teachers' classroom summative assessments and grading by teachers. The word 'intervention' is used to describe the assessment practices studied. In many cases these were 'naturalistic interventions' in the sense that they were part of the ongoing experience of students and not introduced by researchers in order to assess their impact. National Curriculum Tests and similar required assessments were regarded as naturalistic interventions in this respect. Experimental conditions were also included, but, although more controlled, their relevance to normal classrooms needed to be made explicit.

The review also attempts to appraise the weight of evidence provided by the studies, a judgement which includes the methodological soundness, as far as can be judged from the evidence available in the publications reviewed, the relevance of the study type to the review and the appropriateness of the choice of intervention and outcome measures to the questions being researched.

2.2 Review questions

The overall question addressed in the review was:

- What is the evidence of the impact of summative assessment and testing on students' motivation for learning?

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

- How does any impact vary with the characteristics of the students and the conditions of the assessment or testing?
- In those studies where impact on students has been reported, what is the evidence of impact on teachers and teaching?
- What actions in what circumstances would increase the positive and decrease the negative impact on students of summative testing and assessment programmes? In particular, what is the evidence that any impact is increased by 'raising' the stakes?
- What are the implications for assessment policy and practice of these findings?

3. IDENTIFYING AND DESCRIBING STUDIES: METHODS

The review set out to identify as many research studies as possible, of all types, relating to naturalistic and experimental interventions and concerning the processes and outcomes relevant to answering the review question. It involved a process of subjecting studies to successive rounds of close scrutiny and application of criteria until those studies most relevant to the review question were identified. These studies were then assessed for methodological quality and validity as part of the process of data extraction. Thus, following the definition of questions and the development of the protocol there were seven stages to the review: searching for studies; applying criteria to abstracts; obtaining full texts, keywording and summarizing the characteristics of studies; applying refined criteria for final selection; extracting data and evaluating weight of evidence; synthesizing and consultation. The flowchart in Figure 1 represents these stages and each is now briefly described.

3.1 Applying inclusion and exclusion criteria

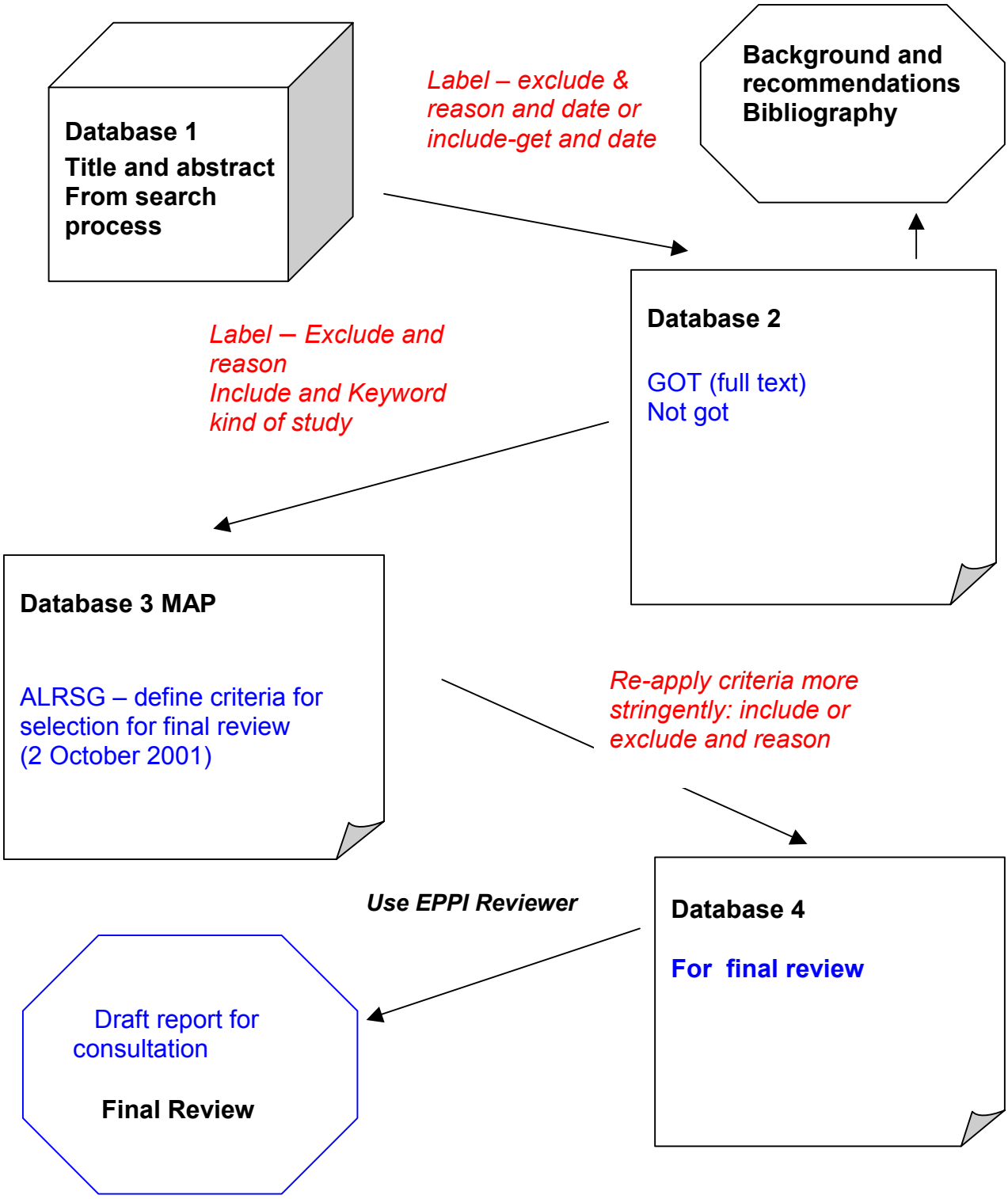
Before obtaining the full text of the identified studies, inclusion and exclusion criteria were applied to the abstracts from electronic databases such as ERIC and BEI that were put into a database for this review. The inclusion and exclusion criteria were as follows:

- *Language of report:* Included studies were written in English. This inclusion criteria was due to the researchers' limited command of other languages. Future reviews may be able to include studies written in other languages.

- *Type of assessment and testing programme:* Studies were included only if they reported on programmes of summative testing or assessment. As noted in the background to this review, previous work has been concerned with assessment for formative purposes and the aims of this review are to consider the evidence in relation to assessment for summative purposes. Studies dealing only with formative assessment were excluded.

- *Type of study and study design:* Studies were included if they evaluated the processes involved in assessment and testing programmes (process evaluations) or the impact of assessment and testing programmes on motivation for learning, (outcome evaluations). Outcome evaluations of a range of study designs and descriptive studies were included. For the purpose of setting out the background and context of the review, a wider range of studies, such as commentaries on existing research, and discussion of relevant issues, was initially included. At the keywording stage, the commentaries and discussion papers were transferred to a separate database, set aside from single studies used in synthesising findings. The keywords allowed reviews of research to be included but these reviews were later excluded as the process of systematic research review is based on the synthesis of primary studies from the synthesis since the EPPI Reviewer handled only single studies.

Figure 1: Flowchart of studies at each stage of the review



- *Setting and population:* Studies were included if they reported on pupils in school or pre-school, between the ages of 4 and 18. Studies dealing only with students outside this age range, in further or higher education or adult education, were excluded. Studies which reported on teacher-related factors/ experiences/ outcomes were only included if the study also reported on student-related factors/ experiences/ outcomes.

- *Date of research:* As the concern of the review was to report on the effect of current assessment policy and practice, most relevant studies were those reported since about 1990, that is, since the beginning of the period of major policy changes in assessment in England and many other countries. However, since high stakes testing in the US began earlier, and because research into examinations for selection and certification has a longer history, no date limit was applied in the initial search for studies.

3.2 Methods for identifying studies

Studies were identified by searching bibliographic databases and registers of educational research, by handsearching current and back numbers of relevant journals, scanning reference lists of already identified reports, making requests to members of relevant associations and using personal contacts. The details of the search strategy are given in Appendix A. The decision about when to stop searching was taken when the full range of databases and obtainable journals identified as relevant; had been searched to the extent that was feasible and considered worthwhile; for example, when searching earlier years of a journal ceased to produce any relevant studies. It was recognised that no search could be exhaustive.

3.3 Methods for characterising included studies: keywording

For this review, four sets of review-specific keywords were added to the core list of 12 sets of keywords produced by the EPPI-Centre (EPPI-Centre, 2002). The keywords are included in Appendix B. They were applied following guidelines produced by the EPPI-Centre. The Guidelines were helpful but needed careful interpretation.

Keywording was useful in drawing attention to studies not meeting the criteria but which slipped through at earlier stages. For instance, if a study could not be categorised in terms of an assessment form and a motivation outcome (review-specific keywords) it was recoded and excluded. The keywords which were most useful in the next stage of defining the map, were those which designated the type of study, and those which identified a particular type of motivation outcome (for example, self-esteem, locus of control, learning dispositions, etc.).

3.4 Methods for quality assurance

As noted above, the guidelines for applying keywords required some interpretation. To check reliability in applying key words, 30 studies were keyworded by two people, including EPPI team members. Agreement was considerable and differences helped in clarifying the meaning of terms.

4. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

4.1 Number of hits and studies included at each stage

Figure 2 summarises the number of studies at each stage of the review. 183 studies were identified as relevant and entered into an EndNote database (Db1). Note that, since much handsearching was required, Db1 is smaller than it might have been as a result of considering abstracts only; that is, with the full text available, decisions were made about exclusion at an earlier stage than might otherwise have been the case.

104 studies met the inclusion criteria. Of these, 51 of these were not empirical studies but were of sufficient relevance to be placed in a separate database labelled for use in background discussion and possible guidance in relation to recommendations. Ten studies were excluded at this stage, leaving 43 which were placed in a third database, (Db3).

4.1.1 Final selection of studies

The 43 studies in Db3 were used to create a summary of the characteristics of the studies found to meet the inclusion criteria. The summary is presented in Table 1.

Table 1: Characteristics of the 43 Studies in Database 3 (Db1)

Source of references	No.	Type of study	No.	Country of study	
Electronic database	6	A. Outcome evaluation		US	27
Personal contact	10	(i) RCT	4	UK	10
Handsearch of journals	14	(ii) Trial	3	Australia	2
Journals online	2	(iii) Pre and post test	0	Canada	1
References from other studies	11	(iv) Post test	1	Israel	1
		(v) Reversal design	0	Morocco	1
		(vi) Cohort study	6	New Zealand	1
		(vii) Case control study	1		
		(viii) Other design	5		
		B. Process evaluation	11		
		C. Economic evaluation	0		
		D. Intervention description	0		
		E. Methods			
		(i) Instrument design	0		
		(ii) Other	0		
		F. Needs assessment	0		
		G. Review			
		(i) Systematic	1		
		(ii) Non-systematic	7		
		(iii) Meta-analysis	1		
		H. Descriptive study	3		

Table 1: Characteristics of the 43 studies in Db 3 (cont'd)

Age range*	No.	Types of learning motivation*	No.	Gender	No.
Primary children	25	i. Intrinsic Motivation or Extrinsic Motivation or Achievement Motivation or Performance learning or Mastery learning	25	Mixed	43
Secondary children	24	ii. Self-Esteem / Self-Efficacy	23		
Post compulsory learners (17- 20)	3	iii. Locus of control/Executive control/ Self-regulated learning	13		
		iv. Learning to learn/learning profile/learning journey	3		
		v. Test Anxiety/stress/phobia	12		
		vi. Learning for meaning	2		
		vii. Learning dispositions	4		

* These categories were not mutually exclusive

The summary, accompanied by a brief description of each study in Db3, was discussed at a meeting of all members of the Review Group and the position of particular studies in the final selection of studies was reviewed. Eight of the included studies were reviews of research. These had been retained at the keywording stage, but were not included in the synthesis which is based on primary research studies. However, these studies have been used to identify specific studies and to inform the background discussion. Sixteen other studies were excluded for various reasons. In some cases they concerned education beyond the age of 19. In other cases, decisions were made to exclude studies where interventions could be regarded as involving assessment but were not specific assessment or test methods: for example, setting or streaming, and studies where the dependent variable was students' achievement rather than motivation. At this stage, 18 studies remained, but a further important study was obtained later and this was included in the final database, giving a total of 19 included studies.

4.2 Characteristics of included studies

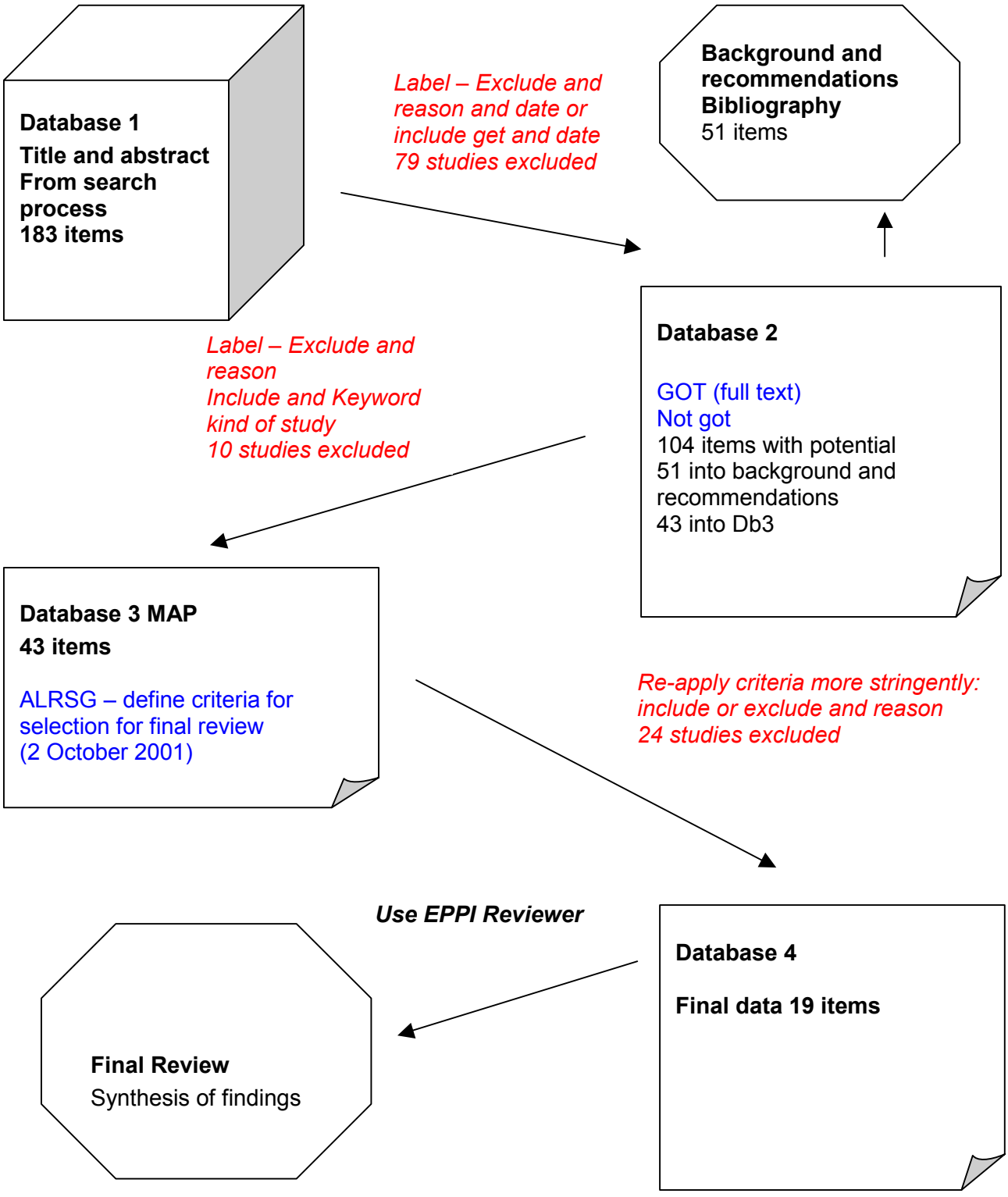
The sources of the included studies are indicated in Table 2.

Table 2 : Sources of the 19 selected studies for Database 4

Source of location of study	Number of studies (N=19)
Database	
BEI	3
ERIC	1
Personal contacts	5
Handsearch of journals	5
References from other studies	5

Table 3 gives the characteristics of the studies in terms of design type, type of intervention, type of outcome reported, country of origin and quality rating. The type of outcomes labelled in this table are derived from a close reading of the texts.

Figure 2: Numbers of items identified at each stage of the review process



This summary of the features of the 19 included studies indicates the ways in which each study was categorised by the keywording process. The keywords deemed to designate aspects of motivation for learning enabled the pattern of three over-arching themes to emerge. These are:

- *What I feel and think about myself as a learner:* self-esteem, self as a learner, learning disposition, attitude to tests, test anxiety
- *The energy I have for the task:* effort; interest, self-regulation
- *How I perceive my capacity to undertake the task:* self-efficacy

Table 3: Features of the 19 selected studies

Design types	**Type of intervention**	**Type of outcome** (> 1 per study)	**Country of origin**
Outcome:	National Curriculum	Effort 9	Canada 1
RCT 3	tests: 4	Self-efficacy 4	Israel 1
Case control 3	State tests: 3	Self-esteem 7	Morocco 1
Post-tests 3	11+ (Northern	Interest 3	UK 8
Other design 4	Ireland) 2	Attitude to tests 5	USA 8
Process: 3	Classroom:	Test anxiety 3	
Descriptive: 3	assessment: 5	Learning	
	Experimental 2	disposition 3	
	Feedback 1	Self-regulation 2	
	16+ exams 2	Self as a learner 4	

5. IN-DEPTH REVIEW: METHODS

5.1 Moving from broad characterisation to in-depth review

The next step after mapping was a close reading and review of the full texts in order to extract the detailed data necessary to answer the review question and to assess the weight of evidence from contributing studies. The process is described in the following sections.

5.2 Methods for extracting data and evaluating weight to be given to evidence

Data extraction was carried out using the*Guidelines for Extracting Data and Assessing Quality of Primary Studies in Educational Research, Version 0.9.4* (EPPI-Centre, 2001) - online version is called EPPI Reviewer. This involved answering 130 to 150 questions (depending on the type of study) about the research reported in each study. The EPPI Reviewer was available for use both online and offline. The process of extracting data from a study could take from four to six hours, depending on the length and complexity of the report. Data were extracted for each study independently by at least two people and difference reconciled before the final entries were made into the EPPI Reviewer.

5.2.1 Evaluation of weight to be given to evidence from each include study

Whilst all the 19 studies met the inclusion criteria and could be characterised using the general and specific keywords, they varied in design, methodology, instruments used and close relevance to the review questions. In order to ensure that conclusions were based of the most sound and relevant evidence, judgements were made about three aspects of each study and these were combined to give an overall judgement of the weight that could be attached to the evidence from a particular study to answer the review question.

The three aspects are outlined in A to D below.

A soundness of methodology

These judgments of how well the study had been carried out were informed by the responses to the following eight aspects of the study in the section of the EPPI Reviewer relating to quality of the study:

- Is the context of the study adequately described?
- Is there sufficient justification for why the study was done in the way it was done?
- Are the aims of the study clear?
- Is there an adequate description of the sample used in the study and how the sample was recruited?

- Is there an adequate description of the methods used in the study, including methods used to collect data and methods of data analysis?
- Have sufficient attempts been made to establish the reliability and/or validity of data collection tools?
- Have sufficient attempts been made to establish the reliability and/or validity of data analysis.
- Are sufficient original data included so that it is possible to mediate between data and interpretation?

In turn, the answers to these questions were informed by data about, for example, the methods for ensuring the validity of data analysis identified in an earlier part of the data extraction process. 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, just that it was not reported. Studies were rated as high, medium or low in relation to methodological soundness according to the number of these criteria that were met (6-8, high; 4-5, medium; 0-3 low).

B Appropriateness of study type and 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 to the focus of the review than others. This was not a judgement on 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. Studies were rated high, medium and low in relation to this aspect.

C Relevance of the topic 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 of the study, that is, how well the nature of the data collected helped to answer the reviews questions. Again the judgements were in terms of high, medium or low relevance.

D Overall weight that can be give 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 favoured C, the relevance of the topic focus. This was justified in terms of the importance of the validity of the data. That is, the studies of greatest weight were those reporting findings relevant to students in school as opposed to, for example, performance on artificial tasks whose relationship to school work was doubtful. This means that if a study was rated as ‘medium’ for C and high for A and B, the overall rating was ‘medium’. A study rated ‘high’ for C and medium for A would be rated ‘high’ overall since failure to meet some of the methodological criteria often reflects the need to report succinctly rather than a flaw in the processes carried out.

5.3 Methods for synthesising findings

Lengthy consideration was given to the various ways in which the findings of different studies could be brought together to form conclusions. In this review of the impact of summative assessment on motivation, the research question sets up summative assessment and testing (the naturalistic or experimental intervention) as the independent variable and motivation for learning as the dependent variable. However, there is no unique dependent variable which can be measured as an outcome, since, as discussed earlier, motivation is a conglomerate of a range of variables. Nor is summative assessment the only factor affecting this complex overarching concept. A simplified view of the relationship is attempted in Figure 3.

The initial search terms that were chosen to identify studies dealing with motivation for learning were broad in scope in order to pick up a range of variables likely to be relevant to the energy or motivation for classroom learning. (See Appendix A: Search Strategy.)

None of the studies dealt with all the variables included in the concept of motivation for learning but they could be grouped according to the particular outcomes that were investigated. These outcomes fell into three distinct and over-arching variables that were found to be integral to motivation for learning. Expressed from a learner's perspective these are:

- *What I feel and think about myself as a learner*
- *The energy I have for the task*
- *How I perceive my capacity to undertake the task*

Thus the task of synthesising the studies, to answer the main review question was tackled through focusing on the impact of summative assessment and tests on students' motivation for learning, examined through these three over-arching themes which are deemed to be integral to it.

5.4 Consultation

The final phase of the methodology was to present the findings in progress to a peer group drawn together by the Assessment and Learning Research Synthesis Group (ALRSG). This conference included 45 experts, representing teacher practitioners (4), Local Authority or independent advisors (7), Government or government agency representatives (11), teacher educators (8) and academics with research interests in assessment (6) and policy (9). A draft copy of the review was sent to all participants before the conference, and the methodology and findings were presented in detail during the conference. There were no significant problems or concerns expressed relating to the methodology, nor to the theoretical framework utilised to analyse the findings. In the second part of the conference, the participants contributed to an exploration of the implications of the findings for policy and practice.

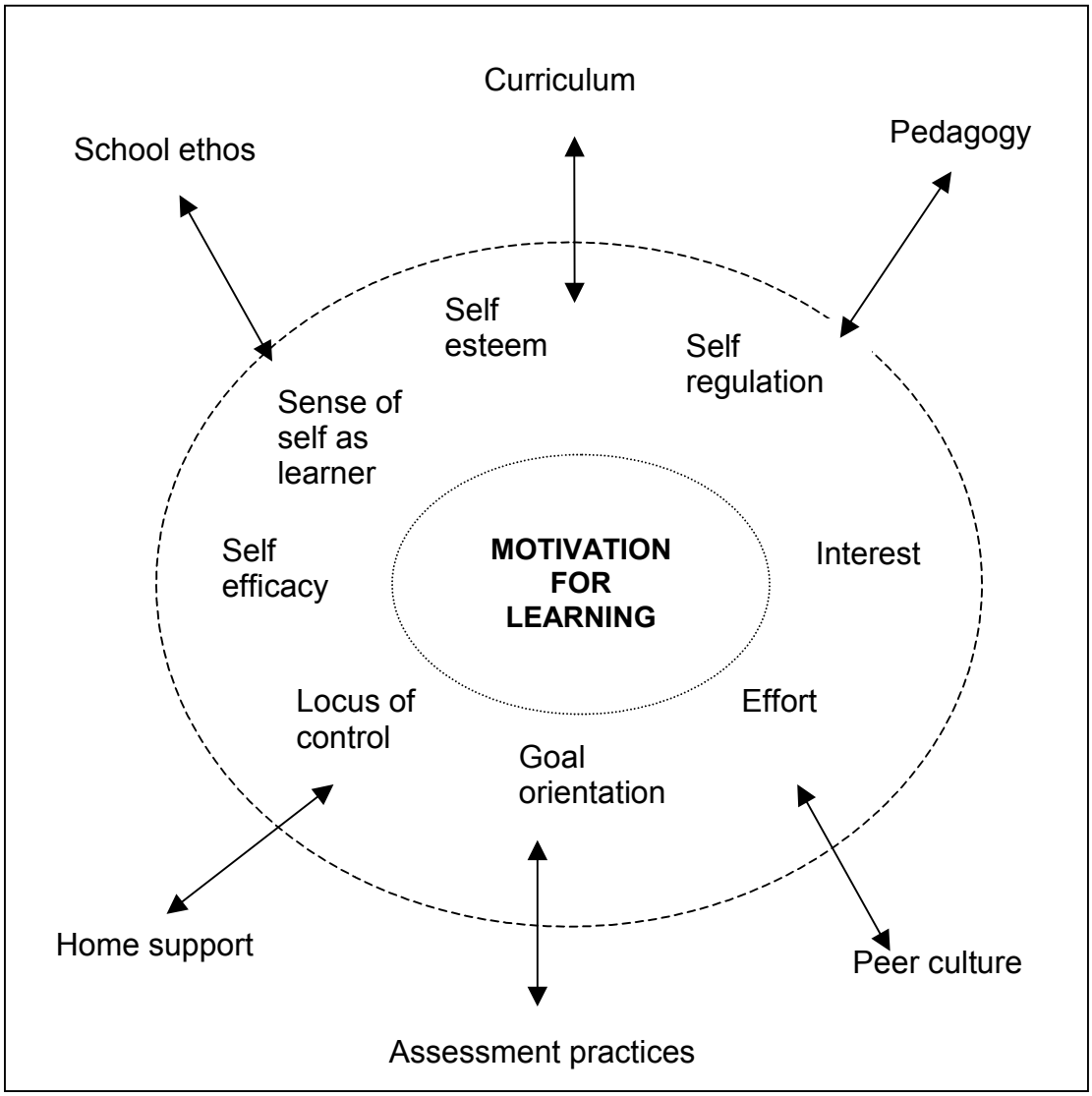
The outcomes of the conference deliberations were recorded and are presented in Appendix D.

5.5 Methods for quality assurance

Data extraction was carried out for all 19 studies independently by two people and for 7 studies by three people, including EPPI-Centre staff. The main difference stemmed from identification of study type, which has implications for the selection of questions used in extracting some of the data. There were some differences also in the detail recorded in completing answers, due in the most part to the inexperience of the reviewers meaning that they could not anticipate what was needed in the subsequent stages of synthesis.

The judgements related to the assessment of weight were made by two researchers working together and agreeing the rating for A, B and C for each study and the basis for arriving at an overall weight.

Figure 3: Some of the variables relating to motivation and factors affecting them



6. IN-DEPTH REVIEW: RESULTS

6.1 Description of included studies

Appendix C provides a structured summary of each of the 19 included studies, giving the type of intervention, study type, outcomes reported, weight given to the evidence and the theme(s) to which the study relates.

The judgements in relation to soundness, appropriateness of study type and design and relevance of the topic focus, and the overall judgement of weight are given in Table 4. It shows that 12 of the 19 studies were identified as having high overall weighting, six as having medium weighting and one as having low weighting.

Table 4: Results of the assessment of weight of evidence for each study

Study	Criteria for methodological quality (max=8)	A Soundness of method	B Appropriatness of study type and design	C Relevance of topic focus to review	D Overall weight of evidence
Benmansour (1999)	4	Medium	High	High	High
Brookhart and DeVoge (1999)	8	High	High	High	High
Butler (1988)	8	High	High	High	High
Davies and Brember (1998)	7	High	High	High	High
Davies and Brember (1999)	4	Medium	High	High	High
Duckworth *et al*. (1986)	8	High	High	High	High
Evans and Engelberg (1988)	6	High	High	High	High
Ferguson and Francis (1979	6	High	High	Medium	Medium
Gordon and Reese (1997)	4	Medium	High	Medium	Medium
Hughes *et al*. (1986)	3	Low	High	Low	Low
Johnston and McClune (2000)	7	High	High	High	High

Study	Criteria for methodological quality (max=8)	A Soundness of method	B Appropriatness of study type and design	C Relevance of topic focus to review	D Overall weight of evidence
Leonard and Davey (2001)	6	High	High	High	High
Little (1994)	3	Low	High	Medium	Medium
Paris *et al.* (1991)	2	Low	High	Medium	Medium
Perry (1998)	5	Medium	High	Medium	Medium
Pollard *et al.* (2000)	6	High	High	High	High
Reay and Wiliam (1999)	4	Medium	High	High	High
Roderick and Engel (2001)	5	Medium	High	Medium	Medium
Schunk (1986)	7	High	High	High	High

Table 5 shows the relationship of the studies to the three themes used in the synthesis for the main review question. It also lists the type of intervention and overall weight of evidence. There were 10 studies (eight of high weight and two of medium weight) relating to *What I feel and think about myself as a learner*. Nine studies (four of high weight, four of medium weight and one of low weight) were related to *The energy I have for the task*. Five studies, all of high weight of evidence, related to *How I perceive my capacity to undertake the task*.

Table 5: Type of intervention, weight of evidence and theme for the included studies

Study	Overall weight of evidence	Type of intervention	*What I feel and think about myself as a learner*	*The energy I have for the task*	*How I perceive my capacity to undertake the task*
Benmansour (1999)	H	Naturalistic	X	X	X
Brookhart and DeVoge (1999)	H	Naturalistic		X	X
Butler (1988)	H	Experimental		X	
Davies and Brember (1998)	H	Naturalistic	X		
Davies and Brember (1999)	H	Naturalistic	X		
Duckworth *et al.* (1986)	H	Naturalistic		X	X
Evans and Engelberg	H	Naturalistic	X		

Study	Overall weight of evidence	Type of intervention	*What I feel and think about myself as a learner*	*The energy I have for the task*	*How I perceive my capacity to undertake the task*
(1988)					
Ferguson and Francis (1979)	M	Naturalistic		X	
Gordon and Reese (1997)	M	Naturalistic	X		
Hughes *et al.* (1986)	L	Experimental		X	
Johnston and McClune (2000)	H	Naturalistic	X		X
Leonard and Davey (2001)	H	Naturalistic	X		
Little (1994)	M	Naturalistic		X	
Paris *et al.* (1991)	M	Naturalistic	X		
Perry (1998)	M	Naturalistic		X	
Pollard *et al.* (2000)	H	Naturalistic	X		
Reay and Wiliam (1999)	H	Naturalistic	X		
Roderick and Engel (2001)	M	Naturalistic		X	
Schunk (1986)	H	Experimental			X

6.2 Synthesis across studies: overall review question

What is the evidence of the impact of summative assessment and testing on students' motivation for learning?

Structure of the synthesis

For the purpose of this review motivation for learning was understood to be a form of 'energy' which is experienced by learners and which drives their capacity to learn, adapt and change in response to internal and external stimuli. Energy for learning is influenced by a potentially large range of variables. These include physiological, affective, conative, cognitive, social, cultural and technical factors. Energy for learning is anthropologically prior to these variables and is a natural phenomenon for human beings, rather like

breathing. Evidence of the presence of energy for learning can be identified in variables that are presumed to be integral to it: self-esteem, locus of control, goal orientation, self-efficacy, etc.

As noted in section 5.3, none of the studies dealt with all the variables integral to motivation. The studies were therefore grouped according to the aspects of motivation they included. The groups were identified as follows:

- *What I feel and think about myself as a learner*: related to self-esteem, self-concept, sense of self as a learner, attitude to assessment, test anxiety, learning disposition
- *The energy I have for the task*: related to effort, interest in and attitude to subject, self-regulation
- *How I perceive my capacity to undertake the task*: related to locus of control, goal orientation, self-efficacy

Thus the synthesis of the review findings in relation to the main review question is presented in terms of the studies which provide data in relation to these three variables. The relationship is also indicated on the evidence tables in Appendix C, in the shaded heading to each study. In three cases, a study relates to two of these variables and in one case to all three.

6.2.1 What I feel and think about myself as a learner

The findings of ten studies were relevant to this theme. Eight of these were rated as having a high weight of evidence and two of medium weight. Table 6 summaries the outcomes reported relevant to the review question, the type of intervention, age group studied and the evidence weight.

Table 6: Studies relevant to the theme *What I feel and think about myself as a learner*

Study	Relevant outcomes reported	Type of intervention	Age group	Country	Overall weight of evidence
Benmansour (1999)	Effort, self-efficacy, test anxiety, learning strategy	Naturalistic	High school	Morocco	H
Davies and Brember (1998)	Self-esteem	Naturalistic	Yrs 2 and 6	England	H
Davies and Brember (1999)	Self-esteem	Naturalistic	Yrs 2 and 6	England	H
Evans and Engelberg (1988)	Attitude to grades, attribution	Naturalistic	Grades 4 to 11	USA	H
Gordon and Reese (1997)	Self-esteem, test anxiety	Naturalistic	Elem. and high school teachers	USA	M
Johnston and	Learning dispositions,	Naturalistic	Yrs 6 and 7	Northern Ireland	H

Study	Relevant outcomes reported	Type of intervention	Age group	Country	Overall weight of evidence
McClune (2000)	locus of control, self-esteem		(age 11)		
Leonard and Davey (2001)	Attitude to tests, self-esteem, test anxiety	Naturalistic	Year 6 (age 11)	Northern Ireland	H
Paris *et al.* (1991)	Self-esteem, attitudes to assessment	Naturalistic	Grades 2-11	USA	M
Pollard et al (2000)	Attitude to tests, learning dispositions	Naturalistic	Years 1-6 (age 5 to 11)	England	H
Reay and Wiliam (1999)	Self-esteem, attitude to tests, test anxiety, self-perception as learner	Naturalistic	Year 6 (age 11)	England	H

Self-esteem

Seven studies dealt directly with measures or perceptions of students' self-esteem and views of themselves as learners. Two reports by Davies and Brember (1998 and 1999) described results of an eight-year study of primary children in England. Using the Lawseq questionnaire as a measure of self-esteem, they followed changes in the self-esteem of successive cohorts of year 2 and year 6 students over a period of eight years, starting two years before the National CurriculumTests were introduced at Key Stage 1. They found a drop in self-esteem for year 2 students, year by year for the first four years, with the greatest change coinciding with the introduction of the National Curriculum Tests. However there was a recovery for later cohorts such that the final, eighth cohort had a higher level of self-esteem than any previous cohort. For year 6 cohorts, there was a rise in self-esteem from year to year with no dip. The self-esteem in year 6 of the students who were tested at year 2 showed little change.

The authors' discussion focused on the drop in self-esteem coinciding with the introduction of KS1 tests. They pointed out that these tests, when first introduced, were cumbersome, lengthy and demanded a great deal of teacher attention to individual students, leaving others feeling undervalued. Moreover, teachers were reeling from the wide-ranging changes taking place, not only in the assessment and curriculum but in school management, relations with parents and various accountability measures. They suggest that the teachers' low morale could have been communicated to the students in addition to the teachers paying less attention to the class as a whole. Once the National Curriculum Tests were simplified and teachers settled to a new regime, the KS1 students' self-esteem rose. For the year 6 students, the tests did not begin until four years after the first KS1 tests and there was time for 'an assessment culture' to have developed in the schools.

More indicative of a long-term impact of the National Curriculum Tests was Davies and Brember's finding that for pre-National Curriculum Test cohorts there was no correlation between self-esteem and achievement as measured by standardised tests in mathematics and reading. Post-National Curriculum Testing, however, there was a small but statistically significant correlation between self-esteem and achievement. This suggests that before the tests were introduced, low-achieving students were no more likely to have low self-esteem than high-achieving students. But after the introduction of the National Curriculum Tests, the low-achievers had a lower self-esteem than their higher achieving classmates. There is, of course, no basis for suggesting that the National Curriculum Tests were a direct cause of the change in correlations; indeed the impact of testing is rarely direct but mediated through a variety of circumstances and people influencing children's affective responses to tests. However, this was a study providing high-weight evidence and it does point to the introduction of the tests as the main factor which differed for the cohorts of students concerned, whatever the mechanism of its impact.

The Northern Ireland 11+ examination was the subject of two of the studies. Johnston and McLune (2000) investigated the impact on teachers, students and students' learning processes in science lessons through interviews, questionnaires and classroom observations. Leonard and Davey (2001) reported the students' perspectives of the process of preparing for, taking and coming to terms with the results of the 11+.

Johnston and McLune used several instruments to measure students' learning dispositions, self-esteem, locus of control and attitude to science and related these to the transfer grades obtained by the students in the 11+ examination. The measures were the Learning Combination Inventory (Johnston, 1996), the B/G steem scale for primary pupils (Maines, 1996) and the Locus of Control Scale for Students (Norwicki, 1973). From the Learning Combination Inventory, they found four main learning dispositions:

- *precise processing:* preference for gathering, processing and utilising lots of data, which gives rise to asking and answering many questions and a preference for demonstrating learning through writing answers and factual reports
- *sequential processing:* preference for clear and explicit directions in approaching learning tasks
- *technical processing*: preference for hands on experience and problem solving tasks; willingness to take risks, to be creative
- *confluent processing:* typical of creative and imaginative thinkers, who think in terms of connections and links between ideas and phenomena and like to see the 'bigger picture'

Classroom observation showed that teachers were teaching in ways that give priority to sequential processing and link success and ability in science to precise/sequential processing. The statistical analysis showed a positive correlation between precise/sequential learning dispositions and self-esteem. The more positive a student's disposition towards precise sequential or technical processing, the higher their self-esteem and the more internal their locus of control. Conversely, the more confluent the pupils' learning orientation, the more external their locus of control and the lower their self-esteem. Interviews with teachers indicated that they felt the need to teach through highly structured activities and transmission of information on account of the nature of the selection tests. However, the learning dispositions of

students showed a preference for technical processing, that is, through first-hand exploration and problem-solving. Thus teachers may be valuing precise/sequential processing approaches to learning more than other approaches and, in so doing, may discriminate against and demoralise students whose preference is to learn in other ways.

The study by Leonard and Davey, funded by Save the Children, was specifically designed to reveal and publish students' views on the 11+. Students were interviewed in focus groups on three occasions, and they wrote stories and drew pictures about their experiences and feelings. The interviews took place just after taking the test, then in the week before the results were announced and finally a week after the results were known. Thus the various phases of the process could be studied at times when they were uppermost in the students' minds. As well as extreme test anxiety, to which we return later, the impact on the self-esteem of those who did not meet their own or others' expectations was often devastating. Despite efforts by teachers to avoid value judgements being made on the basis of grades achieved, it was clear that, among the students, those who achieved grade A were perceived as smart and grade D students were perceived as stupid. The self-esteem of those receiving a grade D plummeted.

Similar perceptions of self-worth resulting from tests were reported by Reay and Wiliam (1999) from a small-scale study of a year 6 class in a London primary school in the term before the Key Stage 2 National Curriculum Tests were taken. Students were interviewed individually and in groups, and extensive classroom observations were made. The data, in the form of quotations and observations, conveyed a class climate in which the tests became the rationale for all that was done and the criterion by which students were judged and judged themselves. As the time for the tests approached, the students began to refer to the levels they expected to achieve. Repeated practice tests showed some students all too well aware of what they could achieve and this led to very low views of their own capabilities. For example: 'For Hannah what constitutes success is correct spelling and knowing your times table. She is an accomplished writer, a gifted dancer and artist and good at problem solving yet none of those skills make her a somebody in her own eyes. Instead she constructs herself as a failure, an academic non person, by a metonymic shift in which she comes to see herself entirely in terms of the level to which her performance in the SATs (sic) is ascribed' (Reay and Wiliam, 1999, p. 346).

Although less weight can be ascribed to the findings of Gordon and Rees (1997) and Paris *et al.* (1991) than to the studies discussed above, these both report on impact of state mandated tests in the United States on the self-esteem of higher and lower achieving students. The differential impact of testing on low achieving students emerged in Gordon and Rees's exploration of the reactions of teachers in the State of Texas to the Texas Assessment of Academic of Skills (TAAS). Through in-depth interviews, they identified teachers' perceptions of the effects of TAAS on students, teachers and teaching. In relation to the self-esteem of students, a strong theme in the teachers' responses was the lowering of self-esteem of students 'at risk'. In another US study, Paris *et al.* (1991) gathered information about the Michigan State mandated tests. They found that high achievers had more positive self-perceptions than low-achievers.

Attitudes to assessment

All the studies relevant to this component of motivation provided high-weight evidence.

Students experience summative assessment regularly in class and not only when taking external tests. Teachers frequently grade students' regular class work or informal assessment tasks and classroom tests and often give feedback in terms of grades. Sometimes the grading systems are simple and related to clear notions of what is 'correct' and sometimes complex grading criteria are used, combining effort and achievement in relation to expectations for individuals or in relation to expectations for the class. Evans and Engelberg (1988) used a questionnaire to study students' attitudes to and understanding of teachers' grades and how these changed with age, from grade 4 to 11.

In terms of understanding of grades, the authors found, as hypothesised, that older students understood simple grades more than younger ones, but even older students did not understand complex systems of grades. The experience of being given a grade, or label, without knowing what it means seems likely to lead to a feeling of helplessness. In terms of attitudes to grades, not surprisingly, higher-achieving students were more likely to regard grades as fair and to like being graded more than lower-achieving students. This dislike indicates that receiving low grades was an unpleasant experience, giving repeated confirmation of personal value rather than help in making progress. It was found that younger students perceived grades as fair, more than older ones, but they also attached less importance to them. Evans and Engelberg also looked at attribution and found that lower achieving and younger students make more external attributions than higher achieving and older students, who used more ability attributions. This suggests that low-achieving students attempt to protect their self-esteem by attributing their relative failure to external factors.

These findings are echoed in the report of Pollard *et al.* (2000) of part of an extensive study of the impact of the 1988 Education Reform Act in England and Wales, known as the PACE (Primary Assessment, Curriculum and Experience) project. A cohort of students was followed throughout their primary school and data was collected by questionnaire, interview, field notes and structured class observations and students' bubble cartoon completions. By the time the cohort reached year 6 and faced the KS2 tests, national testing was well established in schools and its effect was evident in a number of areas. The authors report an increased focus, from the beginning through the 1990s, on performance outcomes rather than learning processes. Although some students recognised that the tests were to do with judging the teaching they received, others were convinced that they had implications for their future in secondary school. Two thirds of the 54 students interviewed were explicitly aware that the national test results constituted some sort of official judgement of them. 'The sense that the KS2 SATs were a high-stakes activity, and could threaten self-esteem, social status or even lead to some form of stigma, was evidenced in many responses.' (Pollard et al, 2000, p220)

Test anxiety

The reactions of students reported by Leonard and Davey (2001) to the Northern Ireland 11+ tests were similar but all the stronger because of the explicit high stakes for the students' futures. They reported that the majority of students approached the tests with fear and anxiety. The students' drawings

gave evidence of the negative feelings for the whole process; only four out of 193 drawings collected could be interpreted as positive towards the tests. Those confident of passing were likely to be more positive to testing but, as in the PACE study, the initial excitement and novelty of taking practice tests soon wore off. Leonard and Davey found that students across all grade levels tended to be highly critical of the 11+ and wanted it to be abolished. They favoured instead, given that selection was inevitable, continuous assessment by the teacher.

An important finding of *Pollard et al.* (2000) emerged from their classroom observations of teachers' assessment interactions with students. These were intended by teachers to be formative but were interpreted by students as purely summative in purpose. Students realised that whilst effort was encouraged, it was achievement that counted. Indeed in the early 1990s, they found that pupils did interpret class assessment interactions with their teacher as helping them in 'knowing what to do and avoiding doing it wrongly'. But in later years, the students were much less positive about assessment interactions that revealed their weaknesses. They reported anxiety, tension and uncertainty in relation to teachers' assessment. Pollard *et al.* concluded that the anxiety that students felt was arguably a consequence of being exposed to greater risk as performance became more important in the teacher's eyes. The also concluded that students incorporated their teacher's evaluation of them into the construction of their identity as learners. Clearly assessment had a severely reduced role in helping learning and became concerned only with achievement, and there was evidence that students were all too aware of this.

Reay and Wiliam (1999) noted that all the students in the class they observed, except the most able boy, expressed anxiety about failure, with girls more anxious than boys. As in the Northern Ireland study, students also disliked the tests, particularly their narrow focus and did not feel that they could do their best under test conditions.

The association of test anxiety with other characteristics was the subject of Benmansour's (1999) study of high school mathematics students in Morocco. Using questionnaire data, Benmansour found four factors in the measurement of goal orientation and related these to test-anxiety self-efficacy and learning strategies. He found that students with strong orientation to getting good grades had high levels of test anxiety and made greater use of passive rather than active learning strategies. Students with a stronger intrinsic motivation (a desire to learn mathematics out of interest) showed a negative relation with test anxiety and a greater use of active learning strategies. He also found greater levels of test anxiety in girls than boys. Although cause and effect cannot be unravelled by this study, it does suggest that test anxiety is related to the use of passive learning strategies and extrinsic motivation.

Students' sense of self as learners

Four studies already discussed in this chapter, all providing evidence of high weight, describe the impact of assessment on students' perceptions of themselves as learners. As this is such a significant part of motivation to learn, it seems worth bringing these findings together.

The direct measurement of learning dispositions by Johnston and McClune (2000) identified different preferred approaches to learning. They found a considerable preference among learners for working things out for themselves

and for hands-on activities in science rather than the transmission of information which was the style adopted by teachers in science lessons. Thus the majority of students were expected to learn in ways that were not comfortable to them and through which they could not learn as well as they might otherwise. The conflict of styles is likely to lead to students assuming that they are not good learners; whereas, with a flexible and varied approach to teaching, a range of learning styles could be accommodated. The reason for teaching in this way, as noted above, was directly attributed by the teachers to the existence and nature of the 11+ selection tests.

The more direct outcome of the tests on sense of self was evident in the studies of Leonard and Davey (2001) and of Reay and Wiliam (1999). They reported that students' judgements about being smart or stupid were inexorably made on the basis of the 11+ grade or the national curriculum level achieved. These became part of the classroom climate – labels ready to be placed on students when results were announced. Many knew their fate beforehand from practice tests and ceased to strive against the inevitable, writing themselves off as learners. The process was not an easy one, as Pollard *et al.* (2000) report, for some low achievers became dysfunctional and demotivated, some 'denied' the tests and others became disruptive. The students' comments and drawings indicated that they closely identified their sense of themselves as people and learners with the KS2 levels.

Key findings from individual studies relating to this theme

Evidence of high weight indicates that:

- After the introduction of the National Curriculum Tests in England, low-achieving pupils had lower self-esteem than higher achieving pupils, whilst beforehand there was no correlation between self-esteem and achievement.
- When passing tests is high stakes, teachers adopt a teaching style which emphasised transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences.
- High-stakes tests can become the rationale for all that is done in classrooms and permeates teachers' own assessment interactions.
- Repeated practice tests reinforce the low self-image of the lower achieving students.
- Tests can influence teachers' classroom assessment which is interpreted by students as purely summative regardless of the teacher's intentions, possibly as a result of teachers' over-concern with performance rather than process.
- Students are aware of the performance ethos in the classroom and that the tests give only a narrow view of what they can do.
- Students dislike selection and high-stakes tests, show high levels of test anxiety (particularly girls) and prefer other forms of assessment.

Evidence of medium weight indicates that:

- The State mandated tests in the US lower self-esteem for 'at risk' students.

6.2.2 The energy I put into the task

Nine studies were relevant to this outcome. Four provided high-weight evidence, four provided medium-weight evidence and one was judged to have only low weight in relation to the review questions. (This study is not discussed for reasons given in the summary on Appendix C.)

Table 7 summaries the outcomes reported of relevance to the review, the type of intervention, age group studied and the evidence weight.

Table 7: Studies relevant to the theme *The energy I put into the task*

Study	Relevant outcomes reported	Type of intervention	Age group	Country	Overall weight of evidence
Benmansour (1999)	Effort, self-efficacy, test anxiety, learning strategy	Naturalistic	High school	Morocco	H
Brookhart and DeVoge (1999)	Effort, self-efficacy	Naturalistic	3rd grade	USA	H
Butler (1988)	Interest	Experimental	5th and 6th grade	Israel	H
Duckworth *et al.* (1986)	Effort, self-efficacy	Naturalistic	High school	USA	H
Ferguson and Francis (1979)	Attitude to subject	Naturalistic	High school	England	M
Hughes *et al.* (1986)	Effort, interest	Experimental	5th grade	USA	L
Little (1994)	Interest	Naturalistic	High school	England	M
Perry (1998)	Effort	Naturalistic	2nd and 3rd grades	Canada	M
Roderick and Engel (2001)	Effort	Naturalistic	6th and 8th grades	USA	M

Amount of effort

The four studies first discussed provided high weight evidence relating to factors that influence the effort that students put into their work.

The study by Brookhart and DeVoge (1999) tested a theoretical model for interpreting results of assessment events in a limited environment. The model included the following variables: level of perceived task characteristics, perceived self-efficacy, amount of invested mental effort, achievement, and the relations between these. Classroom achievement is conventionally measured by classroom assessments that teachers construct or select for this purpose. These assessments are the basis of students' perceptions as to what it is important to learn and where to direct effort in learning. To explore these relationships, two third-grade language arts classes were studied over four classroom assessment events. A description of the level of perceived task

characteristics, perceived self-efficacy, amount of invested mental effort, achievement, and the relations among these for four events in both classroom environments was sought. Four different classroom assessment events were selected in each class, in consultation with the teachers. For each event, a pre-survey was administered to the whole class to collect perceptions of perceived task characteristics and perceived self-efficacy to do the task. A post-survey was administered after the assessment but before students received feedback, to collect perceptions of amount of invested mental effort. Achievement was noted as the score the teacher assigned for student performance on the assessment (i.e. percentage correct). Before each assessment event, four students were selected, in consultation with the teacher, to be interviewed about their perceptions of the assessment.

A key outcome was the significance of feedback from earlier performance on willingness to invest effort in a particular task. Students obtained feedback directly from their previous performance on similar tasks or from the teacher. Feedback that is judgemental and relates to students' capabilities rather than to the task can influence the effort that students put into future learning. Goal orientation was also found to be linked to effort, greater effort being associated with learning goals, specifying the intended learning, as compared with performance goals, specifying what is to be produced.

Duckworth *et al.* (1986) also studied impact of normal classroom grading procedures but in this case with high school students. Their aim was to understand the relationship between effort, motivation, efficacy and futility in relation to type of teacher feedback so as to inform assessment practice. Questionnaires were administered to a cross-section of students in 69 schools to provide indices of effort, motivation, efficacy and futility. Some of the findings echoed those of Brookhart and DeVoge (1999). In particular, Duckworth *et al.* found students' perceptions of communication, feedback and helpfulness of their teachers to be strongly related to feelings of efficacy of study and effort to study.

Butler (1988) tested hypotheses about feedback and its impact on interest in tasks in a randomised controlled trial. Fifth and sixth grade students in Israel were randomly assigned to three experimental conditions of feedback whilst they undertook a convergent task (constructing words from given letters) and a divergent thinking task. Students were scored on both tasks and were also given an interest questionnaire after each session. The three experimental conditions of feedback were:

- Comments only: feedback consisted of one sentence, which related specifically to the performance of the individual child
- Grades only: based on the scores after conversion to follow a normal distribution with scores ranging from 40 to 99.
- Grades plus comments

For the convergent tasks, high achievers scored higher in comments-only conditions and in grades-only conditions than in grades plus comments. For low achievers, those in comments-only conditions scored more highly than those in grades-only conditions and those in grades-only scored more highly than grades plus comments. Thus both high and low achievers did better with grades-only than grades plus comments. For divergent tasks, those under comments-only conditions scored more highly than under grades-only and grades plus comments conditions and there was no significant difference

between the latter two groups. This was the same for high and low achievers. The interest that high achievers expressed in the tasks was similar for all feedback conditions but low achievers expressed most interest after comments only.

Other research reported here (e.g. Pollard *et al.*, 2000) confirms that interest and effort are related and students will put in effort and practice in tasks that interest them. Thus, Butler's conclusions about feedback can be related to the effort that students will put into tasks. She concludes that promoting task involvement by giving task-related, non-ego-involving feedback may promote the interest and performance of most students.

Another study, providing medium weight evidence, reported the impact of a quite different approach to encouraging effort, by using the threat of consequences of failing tests. It was concerned with the effect of the introduction in 1999 by the Chicago Public Schools (CPS) of a requirement for students in the third, sixth and eighth grades to achieve a minimum cut-off score in reading and mathematics on the Iowa Tests of Basic Skills (ITBS) in order to qualify for the next grade, instead of automatic, social promotion from grade to grade. Roderick and Engel (2001) investigated the impact of this policy on sixth and eighth grade students. Their sample consisted of students at risk of being retained; thus they were already seen as having failed at school. All were Afro-American or Latino and many had language or other difficulties and/or home background problems. Baseline data collection included a student interview (semi-structured), collection of student records, and teacher assessments. The teacher assessments asked teachers to report on a variety of areas of student performance using a Likert scale. Following the baseline interview, students were interviewed a second time immediately after taking the ITBS and once during the summer. Retained students were interviewed twice during their retained year.

Roderick and Engel drew on questions from the base line interviews to code work effort, and as a result students were put into four groups:

- Group 1 (53%) were those who were working harder in school as a result of the intervention. They perceived the policy had altered their experiences in school and attitudes towards learning and led them to increase their effort. They reported greater attention to class work, increased academic press (high expectations) from teachers, greater academic effort in and out of class. A higher proportion of these children were at low and moderate risk of retention.
- Group 2 (9%) were those working harder but outside of school, supported by other adults. Most of these students had supportive parents. They were evenly spread across gender and grades and race.
- Group 3 was the 'worrying but not working' group, comprising 34%. These students seldom related what they were doing in school to preparing for the ITBS. There was a higher proportion of 6th graders, males and Latinos in this group.
- Group 4 comprised four students (4%) who were the most highly skilled in the sample and had already met targets in at least one subject.

Across the groups there were differences in age, gender and race. Eighth-graders worked harder than sixth- graders, males less than females and Latinos more likely to be worrying and not working than Afro-Americans. Striking differences according to school support were noted. A school giving

high support was markedly more successful in terms of student effort than a similar school which gave little support. High support meant creating an environment of social and educational support, working hard to increase students' sense of self-efficacy, focusing on task-centred goals, making goals explicit, using assessment to help pupils succeed and creating cognitive maps. They also adopted a strong sense of responsibility for their students. Low teacher support meant teachers not seeing the target grades as attainable, not translating the need to work harder into meaningful activities, not displaying recognition of change and motivation on the part of students, not making personal connections with students in relation to learning goals.

Effort was found to be related to outcome. Almost all students making an effort passed the test at the required level, whilst only a third of students not making an effort did so. The authors conclude that, although the majority of students responded to the policy, the use of high-stakes testing as a negative incentive means that some students will fail and these will be the most vulnerable. The study presents some difficulty in interpretation since it was the threat of the particular sanctions attached to the tests (which had implications for social status as well as self-esteem) rather than the tests themselves that was the negative incentive. However, an important finding is that schools can, by giving the kind of help described for the supporting school, raise students' achievement. Tests on their own, without this kind of support do not raise achievement.

Interest and enjoyment

Ferguson and Francis (1979) studied modes of examination and motivation of students taking the GCE 'O' level examination in English. At the time of this study, candidates could be entered either for an examination (Mode 1) or for continuous course assessment by teachers (Mode 3). A pilot questionnaire was administered to two groups of 100 'O' level English candidates. Factor analysis identified three factors in the data relating to interest and enjoyment. These were enjoyment of English lessons, enjoyment of English through reading and literature and valuing English as against other subjects. Items from these factors were used to create the final version of the questionnaire, which was then administered to almost 800 students from 16 schools and colleges where both examinations were offered. The findings showed no statistically significant differences between examination modes in relation to attitude to English, enjoyment of English versus other subjects; enjoyment of English lessons; importance of English; value of English literature and lack of appeal of English literature. The findings relating to mode indicate that the most extreme attitudes came from mode 1 boys in school and mode 1 girls in school. Mode 1 boys show a strong preference for other subjects; mode 1 girls were positive in their attitude to English. The only difference between modes which was notable was between mode 1 boys in school and mode 3 boys in school, where mode 3 boys prefer English to other subjects. Although there were some differences in attitude resulting from mode of examination, the over-riding differences result from gender and to a lesser extent place of study.

Self-regulated learning

In a study carried out in Canada, Perry (1998) observed the effect on young children's effort and control over learning in classrooms that differed in features related to self-regulated learning (SRL). The evidence from this study

was rated at having medium weight due to the use of an instrument not entirely appropriate for young children. Students in three classes high in encouragement of SRL were compared with two classes of low SRL. (The high SRL teachers offered complex activities, offered students choices, enabled them to control the amount of challenge, to collaborate with peers and to evaluate their work. The low SRL teachers were more controlling, offered few choices and their assessments of their own work were limited to mechanical features, such as spelling punctuation, etc.). Data were collected by questionnaire and interview from the grade 2 and 3 children and classrooms were observed. Both questionnaire and interview data pointed to the children in the high SRL classrooms having interest in their work and being motivated by this (intrinsic motivation). 'They indicated a task focus when choosing topics or collaborators for their writing and focused on what they had learned about a topic and how their writing had improved when they evaluated their writing products. In contrast, the students in the low SRL classrooms were more focused on their teacher's evaluations of their writing and how many they got right on a particular assignment. Both the high and low achievers in these classes were concerned with getting 'a good mark'.'

The findings of Perry (1998) compare interestingly with those of Pollard *et al.* (2000) that children tend to judge their own work in terms of whether it is neat, correct and completed, following the criteria that they perceive their teachers to be using. What Perry adds to this picture is that these criteria can be changed by deliberate action on the part of the teacher. Benmansour (1999) also notes that emphasising assessment promotes students to embrace extrinsic goals and concludes that 'In order to counterbalance the emphasis placed on grades, teachers need to cultivate in students more intrinsic interest and self-efficacy, which are potentially conducive to the use of effective strategies and better performance'.

Key findings from individual studies relating to this theme

Evidence from high-weight studies indicates the following:

- Effort is related to achievement in learning.
- Teachers have a key role in supporting students to put effort into their learning activities.
- Students use feedback from earlier performance on similar tasks in relation to the effort they invest in further tasks.
- Teacher feedback that focuses on students' capabilities rather than the task can influence the effort they put into further learning.

Evidence of medium weight indicates the following:

- Low-achieving children can improve their achievement with the help of supportive teachers or other adults.
- Interest and effort are encouraged in classrooms which encourage self-regulated learning by providing students with an element of choice, control over challenge and opportunities to work collaboratively.
- Teachers can influence the criteria that students use in self-assessment of their work.

6.2.3 How I perceive my capacity to undertake the task

Five studies had relevance to this relationship, dealing in various ways with self-esteem, self-efficacy and self-regulation of learning. All of these provided high-weight evidence. Table 8 summarises the outcomes reported of relevance to the review, the type of intervention, age group studied and the evidence weight.

Table 8: Studies relevant to the theme *How I perceive my capacity to undertake the task*

Study	Outcomes reported	Type of intervention	Age group	Country	Overall weight of evidence
Benmansour (1999)	Effort, self-efficacy, test anxiety, learning strategy	Naturalistic	High school - aged 14 upward	Morocco	H
Brookhart and DeVoge (1999)	Effort, self-efficacy	Naturalistic	3rd grade (Year 3 equiv.)	USA	H
Duckworth *et al.* (1986)	Effort, self-efficacy	Naturalistic	High school-aged 14 upward	USA	H
Johnston and McClune (2000)	Learning disposition, locus of control	Naturalistic	Years 6 and 7 (age 11)	Northern Ireland	H
Schunk (1996)	Self-regulation, self-efficacy	Experimental	4th grade (Year 4 equiv.)	USA	H

Self-efficacy

Brookhart and DeVoge's (1999) study of the relationship between perceptions of task, self-efficacy, effort and achievement, emphasised the role of feedback from work done earlier on students' feelings of self-efficacy in relation to current tasks of the same kind. Students use judgements made by themselves or the teacher in deciding whether they are capable of undertaking work successfully. However their own judgements (as with Pollard *et al*., 2000) are based on the criteria communicated implicitly or explicitly and used by the teacher. Brookhart and DeVoge (1999) reported that, in general, students who perceive themselves as more efficacious will also tend to report putting more mental effort into similar tasks. However, the amount of effort put in would depend on whether the task was judged to be easy. Thus self-efficacy and effort were not always directly related for all students.

The study by Duckworth *et al.* (1986) of high school students, reported that self-efficacy was strongly related to students' perceptions of the feedback and help received from their teachers. The role of teachers in influencing students'

feelings of efficacy and effort was underlined by the finding that it is related to collegiality (the amount of constructive talk about testing) among teachers. The author considered the general atmosphere of encouragement in the school to be important and that it is possible that the informal culture of expectations built up over the years by teacher remarks and reactions operates independently of the specific practices studied.

Locus of control

The study by Johnston and McClune (2000) of the selection test for secondary schools in Northern Ireland was outlined earlier. As one of a series of studies, it investigated learning disposition (preferences for different approaches to learning), self-esteem and perceived locus of control. The authors concluded that there was a close link between performance in the transfer tests, students' learning disposition, student self-esteem and pupil locus of control. There was also a significant gender difference in learning dispositions.

Students who favoured the more structured 'precise/sequential process' approach to learning had a higher self-esteem than those who favoured a more exploratory and creative way of learning. This was possibly because precise/sequential processing aligned with the teaching approach adopted by the science teachers. Those with other preferences were unable to use their preferred learning style and their self-esteem as learners suffered. The researchers' classroom observations showed that teaching and learning was strongly focused on transmission of factual knowledge, with much less emphasis on experiential learning and conceptual understanding in preparation for the selection tests and teachers felt that they had to teach in this way on account of the nature of the tests. Thus the existence of the tests was creating a classroom climate that had a considerable effect on self-esteem and locus of control.

Goal orientation

Schunk (1996) explored self-regulatory processes among children who were learning mathematics, in two linked studies. In both studies, two groups of students were randomly assigned to work under either a learning goal or a performance goal ethos. In the first study, half of each group worked with self-evaluation and half without. In the second study, all students in each goal condition evaluated their performance. Self-efficacy, motivation and achievement were measured. Students were randomly assigned to the experimental conditions and worked with teachers from outside the school, receiving 45 minutes instruction on manipulating fractions in seven sessions over seven days.

Relevant findings for this review are those relating to goal orientation and self-evaluation. In Study 1, the effect of goal orientation was apparent only when self-evaluation was absent. Children under self-evaluation conditions and under learning-goal ethos with no self-evaluation solved significantly more problems that did those with performance goals and no self-evaluation. Self-evaluation scores for performance goals and for learning goal were not significantly different. It appeared from Study 1 that self-evaluation swamped any effect of goal-orientation, so in Study 2 all students engaged in self-evaluation. With self-evaluation held constant, the results showed significant effects of goal orientation for self-efficacy and for skill. The scores of the group

working towards learning-goals were significantly higher than those of the performance-goals group on both measures.

The study of Benmansour (1999) explored Moroccan students' perceived motivational orientations, self-efficacy, test anxiety and strategy use in mathematics. High-school students studying for the Baccalaureate completed a self-report questionnaire (in Arabic, which is the language of instruction) designed to measure motivational goal orientation, self-efficacy and test anxiety. The study used factor analysis and tests of difference in scores to investigate relations between these characteristics and their variation with sex.

The findings indicated that self-efficacy was related to higher intrinsic goal orientations, lower-test anxiety and use of a wider repertoire of strategies including active ones. In terms of frequency of use of active and passive learning strategies, passive ones were by far more frequently used by all students, but intrinsically motivated students were more likely to use active ones as well as passive ones. Although the generalisability of this study is limited, it points to the conclusion that an emphasis on assessment is related to greater extrinsic goal orientation in students, to a lower level of self-efficacy and to a limited use of effective learning strategies. However, the study does not allow the direction of cause and effect to be decided.

Key findings from individual studies in this theme

There is high-weight evidence of the following:

- Feedback on assessments has an important role in determining further learning. Judgemental feedback may influence students' views of their capability and likelihood of succeeding.
- High-stakes assessment can create a classroom climate in which transmission teaching and highly structured activities predominate and which favour only those students with certain learning dispositions.
- Goal orientations are linked to effort and self-efficacy.
- Teacher collegiality is important and should be encouraged by school management.
- An education system that puts great emphasis on evaluation and selectivity produces students with strong extrinsic orientation towards grades and social status.

6.3 Synthesis across studies: the subsidiary review questions

6.3.1 How does any impact of summative assessment and testing vary with the characteristics of the students and conditions of testing?

Information relating to differential impact relating to age, level of achievement and gender of students and to the conditions that affect impact was extracted from the studies as indicated in Table 9.

Table 9: Relevance of studies to variation of impact with student characteristics and conditions of testing

Study	Overall weight of evidence	Age of students	Level of achievement of students	Gender of students	Conditions of testing
Benmansour (1999)	H			X	X
Brookhart and DeVoge (1999)	H				X
Butler (1988)	H		X		X
Duckworth *et al.* (1986)	H				X
Evans and Engelberg (1988)	H	X	X	X	
Gordon and Reese (1997)	M		X		X
Johnston and McClune (2000)	H			X	
Leonard and Davey (2001)	H		X		X
Little (1994)	M				X
Paris *et al.* (1991)	M	X	X		
Perry (1998)	M				X
Pollard *et al.* (2000)	H	X	X		X
Reay and Wiliam (1999)	H		X	X	X
Roderick and Engel (2001)	M	X	X		X

Age of students

Two studies providing evidence of high weight indicated that reactions to grades, attribution and goal orientation vary with students' age. Evans and Engelberg's (1988) study relating to teachers' classroom marking or grading, showed that older students (that is, aged 11 and above) were more likely to have a better understanding of simple grades than younger ones. They were less likely to report teachers' grades as being fair but attached more importance to them than did younger children. Pollard *et al.* (2000) also found that older students were more likely to attribute relative success to effort and ability, whilst younger ones attributed it to external factors or practice. Older

students were more likely to focus on performance outcomes rather than learning processes.

More cautiously,we conclude from medium-weight studies that, in relation to tests, lower achieving older students were more likely to minimise effort and respond randomly or by guessing than younger ones. Thus tests have progressively less validity for these children (Paris *et al.*, 1991). However, under threat of serious consequences for not reaching a required level, eighth-graders were more likely to work harder than sixth-graders (Roderick and Engel, 2001). There is no evidence of age differences in test-taking strategies (checking, monitoring time, etc.). Indeed it was reported that instead of increasing motivation and test wiseness with increasing age, older students feel more resentment, anxiety, cynicism and mistrust of standardised achievement tests (Paris *et al.*, 1991).

Level of achievement

Studies giving high weight evidence show that high achieving students are generally less affected by grading than low achievers (Pollard *et al.*, 2000; Paris *et al.*, 1991). They have a better understanding of grades and their interest is less influenced by whether they receive grades or comments or both (Butler, 1988). Not surprisingly, high achievers think grades are fair, whilst low achievers think they are influenced by outside factors (Evans and Engelberg, 1988).

Results of tests which are 'high stakes' for individual students, such as the 11+, have been found to have a particularly strong and devastating impact on those who receive low grades (Leonard and Davey, 2001). However, tests which are high stakes for schools rather than students (such as the National CurriculumTests in England and State-mandated tests in the US) have hardly less impact. Students are aware of repeated practice tests and the narrowing of the curriculum and only those confident of success enjoy the tests (Reay and Wiliam, 1999). In taking tests, high achievers are more persistent, use appropriate test-taking strategies and have more positive self-perceptions than low achievers. In other words, they become better at taking tests and so the gap between high and low achievers is wider on this account than might be the case in terms of understanding and skills alone. Moreover low achievers become overwhelmed by assessments and demotivated by constant evidence of their low achievement thus further increasing the gap. A greater emphasis on summative assessment thus brings about increased differentiation (Pollard *et al.*, 2000; Paris *et al.*, 1991).

Medium-weight evidence on the differential impact of testing on low achieving students emerged in two studies of State-mandated tests in the US. Gordon and Rees's (1997) exploration of the reactions of teachers in the State of Texas to the Texas Assessment of Academic of Skills (TAAS) found through in-depth interviews with teachers that in relation to the self-esteem of students, a strong theme was the lowering of self-esteem of students 'at risk'. In another study, Paris *et al.* (1991) gathered information about the Michigan State mandated tests. They found that high achievers had more positive self-perceptions than low-achievers.

Studies providing high-weight evidence confirm the findings from medium-weight evidence studies that low achievers are doubly disadvantaged by summative assessment. Being labelled as failures has an impact not just on

current feelings about their ability to learn but lowers further their already low self-esteem thus reducing the chance of future effort and success. There is evidence of less weight that when low achievers have a high level of support (from school or home) which shows them how to improve, some do escape from this vicious circle (Roderick and Engel, 2001).

Gender

The evidence on gender differences derives from four studies providing high-weight evidence. Differences in learning dispositions of boys and girls were found to have particular importance in classrooms that favour certain approaches to learning. It was found that boys are more likely than girls to prefer hands-on experiences and problem-solving and girls were more likely to prefer 'sequential' processing, that is, to have clear directions to follow (Johnston and McClune, 2000). Thus girls are more likely to have a higher self-esteem in classrooms where the dominant teaching strategy, moulded by the pressure of tests, favours sequential processing.

At the same time girls were reported as expressing more test anxiety than boys (Evans and Engelberg, 1988; Benmansour 1999; Reay and Wiliam, 1999). Girls also make more internal attributions of success or failure than boys, with consequences for their self-esteem. No gender differences were found in relation to understanding grades (Evans and Engelberg, 1988).

Conditions of assessment

The conditions that tend to increase or decrease the negative impact of summative assessment relate to the degree of self-efficacy of students, the extent to which their effort is intrinsically or extrinsically motivated, the encouragement of self-regulation and self-evaluation and the pressure imposed by adults outside the school (Perry, 1998; Pollard *et al.*, 2000; Gordon and Reese, 1997; Reay and Wiliam, 1999; Roderick and Engel, 2001).

The importance of self-efficacy in supporting student effort and achievement is a thread in several studies providing high-weight evidence. Feedback has a central role in this since self-efficacy is judged from performance in previous tasks of the same kind (Brookhart and DeVoge, 1999; Butler, 1988; Duckworth *et al.*, 1986). If students have experienced success in earlier performance they are more likely to feel able to succeed in a new task. Feedback that focuses on the task is associated with greater interest and effort, whereas feedback that is ego-involving rather than task-involving is associated with an orientation to performance goals (Butler, 1988; Brookhart and DeVoge, 1999). Goal-orientation, effort and interest are all interconnected. Students who are performance orientated have less interest in the task *per se* (Benmansour, 1999). Students who are task-involved and motivated by interest in the work are less likely to experience high test anxiety than those motivated by achieving a high grade (Benmansour, 1999).

Duckworth *et al.* (1986) reported that feelings of self-efficacy are influenced by students' perceptions of teachers' communication about test expectations. They also found that teachers' own class-testing practices can help to increase self-efficacy if teachers explain the purpose and expectations of their tests and provide feedback. Further, a school's 'assessment culture' influences students' feelings of self-efficacy and effort. Collegiality – meaning constructive discussion of testing and the development of desirable

assessment practice in the school – has a positive effect, whilst a focus on performance outcomes has a negative effect. Brookhart and DeVoge (1999) also found that the way in which teachers present and treat classroom assessment events affects the way students approach them.

From other studies, we can more tentatively conclude that the degree to which learners are able to regulate their own learning also appears to promote students' interest and focus on the intrinsic features of their work (Perry, 1998). Students who have some control over their work by being given choice and who are encouraged to evaluate their own work value the significant content features of their work rather than whether it was correct or not. In other classrooms, students evaluated their work by reference to surface features, such as whether is was neat, well presented and 'right' (Perry, 1998; Pollard *et al.*, 2000). Thus classrooms that allow more self-regulation promote change in the criteria students use in self-evaluation. In conditions where self-evaluation operates, task- or learning- goals promote self-efficacy and achievement (Perry, 1998). Students would like their point of view to be taken into account in the tests they undertake (Leonard and Davey, 2001; Little, 1994).

There is a strong basis of evidence that community pressure is brought to bear on the school for high scores (Gordon and Reese, 1997; Reay and Wiliam, 1999) when test scores are a source of pride to parents. Similarly, parents bring pressure on their children when the result has consequences for attendance at high social status schools (Leonard and Davey, 2001). For many students this increases their anxiety even though they recognised their parents as being supportive (Leonard and Davey, 2001; Reay and Wiliam, 1999).

Key points relating to this sub-question

There is strong evidence of the following:

- In comparison with younger students, older students (aged 11 and above): focus more on performance outcomes; have better understanding of simple grades, are less likely to regard teachers' grades as fair and attach more importance to grades.
- In comparison with higher achievers, low achievers are more affected by grading, have a poorer understanding of grades and have less appropriate test taking strategies and less positive perceptions of themselves as learners.
- In comparison with boys, girls report higher levels of test anxiety, make more internal attributions but are more likely to have dispositions suited to sequential learning.
- Students evaluate their work all the time and how they do this depends on the classroom assessment climate; they will do it in terms of performance rather than learning in summative assessment dominated classrooms.
- Teacher feedback is task-focused, promoting self-efficacy and task involvement.
- The school's assessment culture influences students' self-efficacy.
- Pressure from parents directly on their children and from the community on the teachers increases test anxiety.

There is less strong evidence of the following:

- Lower achieving older students are more likely to minimise effort and respond randomly to test items unless under serious threat.
- Lower achieving students are disadvantaged by being labelled and the lowering of self-esteem that this brings, reducing effort and the chance of future success.
- The degree to which students are able to be self-regulating influences their intrinsic motivation.

6.3.2 In those studies where impact on students has been reported, what is the evidence of impact on teachers and teaching?

The seven studies reporting impact on teaching and teachers as well as on students are indicated in Table 10.

Table 10: Studies relevant to impact on teachers and teaching

Study	Overall weight of evidence	Type of intervention
Gordon and Reese (1997)	M	Naturalistic
Johnston and McClune (2000)	H	Naturalistic
Leonard and Davey (2001)	H	Naturalistic
Perry (1998)	M	Naturalistic
Pollard *et al.* (2000)	H	Naturalistic
Reay and Wiliam (1999)	H	Naturalistic
Roderick and Engel (2001)	M	Naturalistic

All seven studies pointed to very similar effects of high-stakes summative assessment. Findings from studies providing high-weight evidence indicate that the existence of external tests has a constricting effect on the curriculum and on teaching methods (Johnston and McClune, 2000). There is emphasis in teaching on the content of the tests (invariably focused on reading and mathematics and occasionally on other aspects of language and some aspects of science) and much less attention to other subjects. Areas particularly neglected are those related to creativity and personal and social development (Leonard and Davey, 2001; Gordon and Reese, 1997).

State-mandated tests in the US, as in the case of the National Curriculum tests in England, have higher stakes for the schools and teachers than for individual students. But where the test is used for selection, as in Northern Ireland where the 11+ examination is used to select for secondary schools, the tests are high stakes for the students as well. However, it seems that the impact on teachers and teaching appears to be the same: teachers focus their teaching on what is taught and subject the students to repeated practice tests. Direct teaching on how to pass the tests can be very effective, so much so that Gordon and Reese (1997) concluded that students can pass tests 'even though the students have never learned the concepts on which they are being

tested'. Their study provides medium-weight evidence that, as teachers become more adept at this process, they can even teach students to answer correctly test items intended to measure students' ability to apply, or synthesise, even though the students have not developed application, analysis or synthesis skills.'

When they are accountable for test scores but not for effective teaching, teachers are reported as expending a great deal of time and effort in preparing students for the tests (Pollard *et al.*, 2000). They administer practice tests which take up time from learning as well as serving to confirm for the low achievers their self-perception as poor learners. Many teachers also go further and actively coach students in passing tests rather than spending time helping them to understand what is being tested (Gordon and Reese, 1997; Leonard and Davey, 2001). Thus the scope and depth of learning is seriously undermined. But this also affects the validity of the tests, since they no longer indicated that the students have the knowledge and skill needed to answer the questions correctly.

Even when not teaching directly to the tests, teachers reported changing their approach. They adjusted their teaching in ways they perceived as necessary because of the tests, spending most time in direct instruction and less in providing opportunity for students to learn through enquiry and problem-solving (Johnston and McClune, 2000).

The extent to which these features of the classroom teaching were the results of the tests, rather than of some other condition, was illuminated by evidence from studies which followed the introduction of national testing and by the overwhelming opinion of teachers in systems where testing has become an established part of their professional experience. The study by Pollard *et al.* (2000), covering the introduction of the National CurriculumTests in England, reveals an impact on teachers' own classroom assessment practice, lending support to the claim that summative assessment drives out formative assessment. After the introduction of tests, students regarded assessment interactions with their teachers as wholly summative, whereas prior to the tests the same students had regarded these as helping them to learn. Even though teachers intended their assessment interactions to be formative, the subtle change in their discourse indicated a summative, performance-related approach that was evidently communicated to the students. Such changes could, of course, have been a natural consequence of maturity but although research evidence does support the interpretation that older students take teachers' assessment more seriously and tend to embrace performance goals more than younger children, the change over time is not entirely explained in this way.

Other studies point to a real change in teachers' behaviour (Johnston and McClune, 2000) and also show how readily students pick up from their teacher the signs of what is valued and will gain approval (Pollard *et al.*, 2000). Thus, as teachers become more performance-centred, students pick up the criteria being used and judge their own work accordingly. There is less strong evidence that teachers can influence children's self-assessment to focus on learning processes (for example, Perry, 1998), but students are unlikely to use such criteria whilst their teachers' assessment and teaching methods implicitly, and in some cases explicitly, reflect performance goals.

Whether or not testing changes the support a school provides to the lowest achieving students is not clear. There is medium-weight evidence that schools vary in the support they give (Roderick and Engel, 2001) and that fewer students would give up on themselves as learners if more schools worked to raise these students' sense of self-efficacy, by focusing on task- and learning-centred goals and using assessment to help them succeed. This underlines the importance of formative assessment but at the same time argues for action that prevents the low self-esteem from developing in the first place.

Key findings relating this sub-question

There is high weight evidence that:

- When passing tests is high stakes, teachers adopt a teaching style which emphasises transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences.
- External tests have a constricting effect on the curriculum, resulting in emphasis on subjects tested at the expense of creativity and personal and social development.
- High stakes tests often result in a great deal of time being spent on practice tests and the valuing of test performance and undervaluing of other student achievements.
- Teachers' own assessment becomes summative in function rather than formative

There is medium weight evidence that:

- Teachers can be very effective in training students to pass tests even when the students do not have the understanding or higher order thinking skills that the tests are intended to measure.
- Teachers can influence students' self-assessment criteria towards learning processes.

6.3.3 What actions in what circumstances would increase the positive and decrease the negative impact on students of summative testing and assessment programmes? In particular, what is the evidence that any impact is increased by 'raising the stakes'?

Table 11 gives information about the studies contributing to answering this research sub-question by providing evidence about actions that might decrease the negative and increase the positive impact of summative assessment and about the effect of raising the stakes.

Table 11: Studies providing evidence relating to changing the impact and to raising the stakes

Study	Overall weight of evidence	Decrease the negative impact	Increase the positive impact	Impact of raising the stakes
Benmansour (1999)	H	X		X
Brookhart and DeVoge (1999)	H	X		
Butler (1988)	H	X		
Davies and Brember (1998)	H			X
Davies and Brember (1999)	H			X
Duckworth *et al.* (1986)	H	X	X	
Evans and Engelberg (1988)	H	X		
Gordon and Reese (1997)	M	X		X
Johnston and McClune (2000)	H	X		X
Leonard and Davey (2001)	H	X	X	X
Little (1994)	M		X	
Paris *et al.* (1991)	M	X		
Perry (1998)	M		X	
Pollard *et al.* (2000)	H		X	X
Reay and Wiliam (1999)	H	X		X
Roderick and Engel (2001)	M	X	X	X
Schunk (1996)	H	X	X	

Actions that would decrease the negative impact of summative assessment

There is strong evidence that one way to decrease the negative impact of summative assessment would be to end the practices of focusing teaching on the test content, training students to pass the tests and using class time for repeated practice tests (Johnston and McClune, 2000; Leonard and Davey, 2001; Gordon and Reese, 1997; Reay and Wiliam, 1999; Paris *et al.,* 1991).

Added to this, a study providing medium weight evidence indicates the value of ending the use of high-stakes testing as a negative incentive, as its use may result in some low achieving pupils failing and suffering further lowering of self-esteem (Roderick and Engel, 2001).

The following more positive action is suggested by high-weight evidence:

- Promote learning goal orientation rather than performance orientation (Brookhart and DeVoge, 1999; Roderick and Engel, 2001; Schunk, 1996).
- Cultivate intrinsic interest in the subject and put less emphasis on grades (Benmansour, 1999) but make grading criteria explicit (Evans and Engelberg, 1988)
- Emphasise teaching approaches that encourage collaboration among students and cater for a range of teaching styles (Pollard *et al.*, 2000; Reay and Wiliam, 1999; Johnston and McClune, 2000)
- Explain and acknowledge the reasons for, and the implications of, tests (Leonard and Davey, 2001; Pollard *et al.*, 2000)
- Provide feedback to students about their performance in a form that is non-ego-involving and non-judgemental (Brookhart and DeVoge, 1999; Butler, 1988) and help students to interpret it (Duckworth *et al.*, 1988)
- Broaden the range of information used in assessing the attainment of students (Reay and Wiliam, 1999).

Medium-weight evidence adds the following action:

- Reduce the 'stakes' attached to test results by broadening the base of information used in evaluating the effectiveness of schools (Gordon and Reese, 1997).

Actions that would increase the positive impact of summative assessment

There is a sense in which avoiding the negative impact implies supporting a positive impact. Thus several positive actions can be identified in the list above: for example, in the type of feedback given and the communication to students of reasons and explanations about assessment. However there are some actions indicated in the studies which would enable summative testing and assessment to take a positive role in students' learning.
There is high-weight evidence in favour of the following actions:

- Ensure that the demands of the tests are consistent with the expectations of teachers and the capabilities of the students (Duckworth *et al.,* 1988)
- Involve students in decisions about testing (Little, 1994; Leonard and Davey, 2001).
- Develop students' self-assessment skills and use of learning rather than performance criteria (Pollard *et al.*, 2000; Schunk, 1996).
- Develop a constructive and supportive school ethos in relation to tests (Duckworth *et al.,* 1988).
- Use assessment to convey a sense of learning progress to students (Duckworth *et al.*, 1988; Roderick and Engel, 2001).

There is also medium-weight evidence in favour of the following actions:

- Supporting low-achieving students' self-efficacy by making learning goals explicit and showing them how to direct effort in learning (Roderick and Engel, 2001)
- Creating a classroom environment that promotes self-regulated learning (Perry, 1998)

The impact of 'raising the stakes'

One mechanism by which the 'stakes are raised' for students is the threat of action based on the results, a practice which inevitably produces failure for students who feel that the gap they have to close is too great (Roderick and Engel, 2001). Reay and Wiliam (1999) also note that threats to schools posed by poor National Curriculum Tests results put teachers under pressure to increase scores by whatever means, regardless of the longer term impact on students' learning.

There is high-weight evidence of the following effect of raising the stakes:

- There is an increase in test anxiety (Benmansour, 1999; Leonard and Davey, 2001; Pollard *et al.*, 2000).
- Students feeling anxiety as a consequence of their sense of being exposed to greater risk as their teacher raised the stakes (Pollard *et al.*, 2000).
- The pressure increases on students to do well resulting from the aspirations of parents and teachers (Davies and Brember, 1998; Leonard and Davey, 2001).
- Teaching is focused on the content of the tests and teaching methods confined to transmission modes which favour sequential learning styles (Johnston and McClune, 2000).
- The use of repeated practice tests impresses on students the importance of the tests and leads to students adopting test-taking strategies designed to avoid effort and responsibility and which are detrimental to higher order thinking (Paris *et al.*, 1991; Reay and Wiliam, 1999).

There is strong evidence that these effects are similar in high and low achieving schools (Johnston and McClune, 2000; Pollard *et al.*, 2000) and less firm evidence that they apply equally to high and low achieving students (Gordon and Reese, 1997). Gordon and Reese (1997) conclude that high-stakes testing has negative effects on the curriculum, teacher decision-making, students' learning, school climate, and teacher and students' self concept and motivation.

6.3.4 What are the implications for assessment policy and practice of these findings?

In order to explore the implications of the review as fully as possible, the planned processes of the review included a consultation conference with invited policy-makers and practitioners. This was held to discuss the outcomes of the synthesis of the 19 studies with the express purpose of eliciting the participants' views first on the credibility of the findings and then on the implications for assessment policy and practice.

The draft findings from the review were discussed in the context not just of summative assessment but against the wider background of assessment in education. With this in mind it is perhaps helpful to recall that what makes assessment 'formative' or 'summative' is what happens to information that is gathered about students' attainments and what use is made of it. When the use is 'by learners and their teachers (in order) to decide where the learners are in their learning, where they need to go and how best to get there' (ARG, 2002), the assessment is described as formative or 'assessment *for* learning'.

When the use is to report to students and others where the students have reached in their learning in relation to overall goals and to monitor their progress over time, the assessment is described as summative or assessment *of* learning. As suggested in the Background section of this report, it is important to use assessment for both these purposes and in theory there is no reason why formative assessment and summative assessment should not co-exist in educational practice. Indeed the same information could be used for both purposes, as suggested by Harlen and James (1997). That is, if learning is the aim of education then assessment can help learning through its formative and summative role. However, this will only be the case if the information is concerned with all aspects of students' attainment and not narrowly confined to those achievements that are measured by standardised tests. Here we are taking attainment to mean the full range of achievements across cognitive, conative and affective domains (DfEE, 2000).

The research reviewed in this study points to practices in using assessment information for summative purposes which have the effect of hindering rather than supporting the learning of some, and in some cases, all, students. The negative effects on the *learners* of assessment for summative purposes have been shown by one or more of the studies providing high weight evidence to be the following:

- A lowering of the self-esteem of the less successful students, which can reduce their effort and image of themselves as learners (Davis and Brember, 1998; Johnston and McClune, 2000; Leonard and Davey, 2001; Reay and Wiliam, 1999)
- A shift towards performance goals rather than learning goals, which is associated with less active and less deep learning strategies and with interest in achievement *per se* rather than interest in the subject (Pollard *et al.*, 2000; Schunk, 1996)
- The creation of test anxiety, which differentially affects students (Leonard and Davey, 2001; Benmansour, 1999; Pollard *et al.*, 2000)
- Judgements of value being made about students, by themselves and others, on the basis of achievement in tests rather than their wider personal attainment (Pollard *et al.*, 2000; Reay and Wiliam, 1999; Evans and Engelberg, 1988).
- The restriction of their learning opportunities by teaching which is focused on what is tested and by teaching methods which favour particular approaches to learning (Johnston and McClune, 2000; Reay and Wiliam, 1999; Evans and Engelberg, 1988).

It is important to emphasise the inter-relationships among these aspects of motivation and experience. For example, test anxiety is greater the more students are oriented to performance goals; the more students are tested, the more they identify their goals in terms of test results and the more they perceive themselves in terms of their test performance; lower performance on earlier occasions influences effort and self-efficacy for further tasks.

On the other hand, several studies showed that students' effort and interest can be enhanced by assessment when feedback is in terms of how improvements can be made and focuses students on task, or learning goals. Self-assessment has also been shown to increase learning and self-efficacy particularly when linked with learning goal orientation. The study by Roderick and Engel (2001) provided medium-weight evidence that summative assessment can be used to increase the effort of some of the lowest achieving

students if the required performance is not too far above their current levels *and* if the students are helped to develop their self-efficacy and are given specific task-related learning goals

In turning to consider implications for practice and for policy, the review brought together the views of the conference participants as well as the findings from the studies reported in earlier sections. Evidence relating to a) the circumstances found to be associated with the negative impact on some aspects of motivation for learning and b) the circumstances found to reduce the negative impact (and in some cases to have a positive effect on some aspects of motivation for learning) had been reviewed in relation to the third sub-question. Outcomes of the consultation conference added the following:

- Using test results to set targets for schools rather than identifying learner-centred targets
- Relying too closely on students' test results in evaluating the effectiveness of schools
- Creating performance tables of schools based on students' achievements in tests

The implications for policy and practice are to try to move in directions that are likely to end the practices that have a negative impact and to enhance those that have a positive impact on motivation for learning. Specific recommendations are drawn from these points in Chapter 8.

7. DISCUSSION

7.1 Re-statement of principal findings

7.1.1 Results of data extraction

The main review question was:

- *What is the evidence of the impact of summative assessment and testing on students' motivation for learning?*

In order to report findings relating to this question, studies were grouped according to their relevance to these three themes:

- *What I feel and think about myself as a learner: related to self-esteem, self-concept, sense of self as a learner, attitude to assessment, test anxiety, learning disposition*
- *The energy I have for the task: related to effort, interest in and attitude to subject, self-regulation*
- *How I perceive my capacity to undertake the task: related to locus of control, goal orientation, self-regulation, self-efficacy*

The combined findings relating to these three themes, using the high-weight evidence, was as follows:

- After the introduction of the National CurriculumTests in England, low-achieving pupils had lower self-esteem than higher-achieving pupils, whilst beforehand there was no correlation between self-esteem and achievement.
- When passing tests is high stakes, teachers adopt a teaching style which emphasises transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences.
- Repeated practice tests reinforce the low self-image of the lower achieving students.
- Tests can influence teachers' classroom assessment which can be interpreted by students as purely summative regardless of the teacher's intentions, possibly as a result of teachers' over-concern with performance rather than process.
- Students are aware of a performance ethos in the classroom and that the tests give only a narrow view of what they can do.
- Students dislike high stakes tests, show high levels of test anxiety (particularly girls) and prefer other forms of assessment.
- Teachers have a key role in supporting students to put effort into their learning activities.
- Feedback on assessments has an important role in determining further learning. Students use feedback from earlier performance on similar tasks in relation to the effort they invest in further tasks.

- Teacher feedback that is ego-involving rather than task-involving can influence the effort students put into further learning and their orientation towards performance rather than learning goals.
- High stakes assessment can create a classroom climate in which transmission teaching and highly structured activities predominate favouring only those students with certain learning dispositions.
- High stakes tests can become the rationale for all that is done in classrooms and permeates teachers' own assessment interactions.
- Goal orientations are linked to effort and self-efficacy.
- Teacher collegiality is important in creating an assessment ethos that supports students' feelings of self-efficacy and effort.
- An education system that puts great emphasis on evaluation which produces students with strong extrinsic orientation towards grades and social status.

There is medium-weight evidence that:

- The State mandated tests in the US lower self-esteem for 'at risk' students.
- Low-achieving children can improve their achievement with the help of supportive teachers or other adults.
- Interest and effort are encouraged in classrooms which encourage self-regulated learning by providing students with an element of choice, control over challenge and opportunities to work collaboratively.
- Teachers can influence the criteria that students use in self-assessment of their work.

7.1.2 Subsidiary questions

How does any impact of summative assessment and testing vary with the characteristics of the students and conditions of testing?

There is high-weight evidence of the following:

- In comparison with younger students, older students (age 11 and above) focus more on performance outcomes; have better understanding of simple grades, are less likely to regard teachers' grades as fair and attach more importance to grades.
- In comparison with higher achievers, low achievers are more affected by grading, have a poorer understanding of grades and have less appropriate test taking strategies and less positive perceptions of themselves as learners.
- In comparison with boys, girls report higher levels of test anxiety, make more internal attributions but are more likely to have dispositions suited to sequential learning.
- Students evaluate their work all the time and how they do this depends on the classroom assessment climate; they will do it in terms of performance rather than learning in summative assessment dominated classrooms.
- Teacher feedback that is task focused promoting self-efficacy and task involvement
- The school's assessment culture influences students' self-efficacy.
- Pressure from parents directly on their children and from the community on the teachers increases test anxiety.

There is less strong evidence of the following:

- Lower achieving older students are more likely to minimise effect and respond randomly to test items unless under serious threat.
- Lower achieving students are disadvantaged by being labelled and the lowering of self-esteem that this brings, reducing effort and the chance of future success.
- The degree to which students are able to be self-regulating influences their intrinsic motivation.

In those studies where impact on students has been reported, what is the evidence of impact on teachers and teaching?

- When passing tests is high stakes, teachers adopt a teaching style which emphasises transmission teaching of knowledge, thereby favouring those students who prefer to learn in this way and disadvantaging and lowering the self-esteem of those who prefer more active and creative learning experiences.
- External tests have a constricting effect on the curriculum, resulting in emphasis on subjects tested at the expense of creativity and personal and social development.
- High stakes tests often result in a great deal of time being spent on practice tests and the valuing of test performance and undervaluing of other student achievements.
- Teachers' own assessment becomes summative in function rather than formative.

There is medium-weight evidence that:

- Teachers can be very effective in training students to pass tests even when the students do not have the understanding or higher order thinking skills that the tests are intended to measure.

What actions in what circumstances would increase the positive and decrease the negative impact on students of summative testing and assessment programmes? In particular, what is the evidence that any impact is increased by 'raising the stakes'?

There is high-weight evidence that these actions decrease the negative impact:

- Promoting learning goal orientation rather than performance orientation.
- Cultivating intrinsic interest in the subject and put less emphasis on grades.
- Explaining and acknowledging the reasons for, and the implications of, tests.
- Providing feedback to students about their performance in a form that is non-ego-involving and non-judgemental.
- Broadening the range of information used in assessing the attainment of students.

And there is medium-weight evidence for the following action:

- Reducing the 'stakes' attached to test results by broadening the base of information used in evaluating the effectiveness of schools.

There is high-weight evidence that these actions increase the positive impact:

- Ensuring that the demands of the tests are consistent with the expectations of teachers and the capabilities of the students
- Involving students in decisions about testing
- Developing students' self-assessment skills and use of learning rather than performance criteria
- Developing a constructive and supportive school ethos in relation to tests.
- Using assessment to convey a sense of learning progress to students

And there is medium-weight evidence for the following actions:

- Supporting low-achieving students' self-efficacy by making learning goals explicit and showing them how to direct effort in learning (Roderick and Engel, 2001)
- Creating a classroom environment that promotes self-regulated learning (Perry, 1998)

There is high-weight evidence of the following effect of raising the stakes:

- Increase in test anxiety (Benmansour, 1999; Leonard and Davey, 2001; Pollard *et al.*, 2000).
- Students feeling anxiety as a consequence of their sense of being exposed to greater risk as their teacher raised the stakes (Pollard *et al.*, 2000).
- Increase in the pressure on students to do well resulting from the aspirations of parents and teachers (Davies and Brember, 1998; Leonard and Davey, 2001).
- Teaching being focused on the content of the tests and teaching methods confined to transmission modes which favour sequential learning styles (Johnston and McClune, 2000).
- The use of repeated practice tests which impresses on students the importance of the tests, and leads to students adopting test-taking strategies designed to avoid effort and responsibility and which are detrimental to higher order thinking (Paris *et al.*, 1991; Reay and Wiliam, 1999).

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

A consultation conference with invited policy-makes and practitioners reviewed the above findings and added the following from their experience to the practices that are associated with a negative impact of summative assessment on students and schools:

- Using test results to set targets for schools rather than identifying learner-centred targets
- Relying too closely on students' test results in evaluating the effectiveness of schools
- Creating performance tables of schools based on students' achievements in tests

Specific recommendations drawn from these findings are given in Chapter 8.

7.2 Strengths and weaknesses of the review

7.2.1 Methodology of the review

The key characteristic of this review lies in its description as a 'systematic review', a phrase which needs to be unpacked to identify the particular processes which were seen as strengths and those which are potential weaknesses.

The features that are regarded as strengths come under the headings of planning, searching, recording, selecting, checking (quality assurance) and assessment of weight of evidence from individual studies.

Planning

In relation to planning, the strengths of the methodology lay in the specification of the review questions and the time spent on deciding what, where and how to search for studies. The specification of the review question required a balance between being too general and too specific. This balance is particularly critical in education, where contexts, processes and outcomes are complex. To focus a question too narrowly has several disadvantages in education, despite the obvious potential for identifying relevant studies more precisely. Reducing the question to a specified outcome of a single controllable factor risks, firstly, not finding any studies exactly addressing this question and, secondly, if there are such studies, being unable to relate their findings to the real situation of classroom practice. On the other hand, to have too broad a question means that it is difficult to extract specific evidence from the background of 'noise' in a range of studies which are of relevance to the general debates in the area of the review. In the present review it was found essential to keep the focus on student outcomes relevant to motivation that could be ascribed to the effect of summative assessment. Other students outcomes, such as achievement, were not considered unless motivation was also reported and other impacts of summative assessment, such as on the curriculum and classroom practice were only considered in relation to their mediation of the impact of assessment on student motivation.

Searching

In relation to searching, being systematic meant covering as many sources as possible, and, perhaps more importantly, recording the search process. The state of the art of entering educational research into databases lags far behind that in other disciplines, particularly medicine. Thus it was necessary not only to search all the relevant electronic databases but also to follow up citations in earlier reviews and in obtained papers, to hand search journals held in the library as well as those online (which were very limited at the time of this review) and to use personal contacts. No search can, of course, be comprehensive but a strength of this methodology was in recording details, for example of dates of journals that were hand-searched and procedures for searching data-bases, so that its limitations are made explicit.

Recording

Careful recording was also a key feature in relation to selecting studies found to be relevant to the review questions. That only 19 studies were selected for detailed data extraction from an initial number of 183 found in the search, may seem to be a weakness. Indeed some participants at the consultation conference were critical of what seemed to be 'missing data' residing in the other studies. However, the progressive focusing that is central to the EPPI review procedures included documentation of reasons for including and excluding studies. Thus at three points, studies were labelled according to the criteria used for exclusion: Db1 (based on titles and abstracts); Db2 (based on the full text) and Db3 (based on consultation within the review group). Thus rather than being a weakness, the identification of 19 studies meant that attention was given to the most relevant studies for the purposes of answering the review question and that possible obfuscation of the main issues in a wider range of less relevant studies was avoided.

Quality assurance

A further aspect of being systematic was the checking of decisions through double, and sometimes treble, independent action. The application of key words to studies in Db2 was checked in this way and data extraction of each of the 19 studies was also carried out by at least two reviewers and any differences reconciled before the findings were stored in the EPPI Reviewer database. A practical disadvantage of this checking was the extension of the time required for what was already a length process. For example, the data extraction for one study could take between 4 to 6 hours, depending on length and detail, and this time was effectively trebled by double extraction and reconciliation. Being systematic has considerable time and thus resource implications which are not currently reflected in the funding and status of reviews in education.

Assessment of the weight of evidence from individual studies

The primary studies included in this systematic review had been undertaken for a range of different reasons which might not have been that close to the question of this review. The assessment of weight of evidence in relation to the review questions was a feature that enabled the most relevant and sound studies to be given greater weight in the conclusions of the review and the implications drawn from them. It was necessary to know how confident to be of the findings of studies since the 19 included studies reported a range of different types of outcome. As the evidence for a certain finding might come from one or a small number of studies, the dependability of these studies was clearly a matter of importance.

The role of the Review Group

A strength of the methodology was the establishment of a Review Group comprising experts in educational assessment, educational psychology and potential users of the review outcomes. Members of the Group were variously involved in advising on strategy, consulting on substantive content and taking part in the reviewing process. In practice, however, only the members of the Review Group who were also members of the EPPI-Centre team took part in the reviewing process in addition to the two authors of this report, for reasons of work load. Although this arrangement facilitated close communication and

efficient use of time, it was vulnerable to circumstances such as illness or employment change which could have derailed the completion of the review, although fortunately this did not happen. Also, since all the procedures were new, members of the Review Group recognised that closer involvement would have facilitated their understanding of the details of the review process. However, regular meetings and report on progress enabled the Review Group to fulfil a strategic role and to help in key decisions. Whilst the Review group had no formal input from parents or students, referees commented that this was not of great relevance for this review. The views of students on the review findings were gathered by two practitioner members of the Review Group.

The consultation conference

Since the review undertaken here was specifically aimed at informing policy and practice in education, one of its strengths was the incorporation of a consultation conference as part of the methodology. The details of this conference have been given earlier and its findings are reported in Appendix D. The responses confirming the consistency of the findings with the experience of the practitioners and researchers attending supported the validity of the review outcomes. The conference outcomes also provided a basis for identifying recommendations for policy and practice.

Limitations of the review

It was acknowledged by all involved that the procedures and instruments for the systematic review of educational research were being developed during the time that this review was being conducted. There was inevitably a tension between following the already established EPPI procedures, derived from the field of health education, in order to give them a fair trial, and being creative in adapting them to suit the features of educational research. The major weaknesses that can be identified in the methodology, in addition to those already mentioned, arose either from a mismatch between parts of the instruments and the features of educational research that they were intended to describe, or unfamiliarity of the reviewers with the process as a whole.

Particular difficulties were experienced in identifying types of study in the form required by the EPPI-Centre keywords. The process of summarising the characteristics of research studies was also a possible weakness since, at the stage of identifying keywords, the most appropriate ways of describing components of motivation had not been clearly identified. The reason for this relates to the lack of a consistent theory of motivation in education. A further area of difficulty was accommodating the features of descriptive studies and process evaluation in data extraction.

Following the EPPI guidelines meant that all the studies included in this review were ones in which empirical data were collected. Consequently the treatment of other reviews and of theoretically based papers was less systematic. Given the range of alternative ideologies in education and their impact on research styles and procedures, this is an aspect requiring further attention as the process of systematic review of educational research evolves.

7.3 Relationship with other reviews

Eight articles that were reviews of research relating to assessment and testing were identified as relevant to the current review: McDonald (2001), Madeus and Clarke (1999), McNeil and Valenzuela (2000), Black and Wiliam (1998), Kellaghan, Madaus and Raczek (1996), Crooks (1988), Ames (1992) and Natriello (1987). None of these was a systematic review in the meaning of this term as adopted for EPPI-Centre reviews. Most reported on a wide range of assessment purposes and practices and comments here are restricted to those relevant to the current review question. Their findings in these aspects fully support those reported above.

The most relevant review to the current review questions underpins the discussion of **Kellaghan, Madaus and Raczek** (1996) in the AERA monograph *The Use of External Examinations to Improve Student Motivation*. Significantly, one of their conclusions was that too little account is taken of the complexity of the factors relating to motivation. The interaction of different aspects of motivation with a variety of personal characteristics means that what motivates some students may alienate others. They placed considerable emphasis on the goal orientation of students. They concluded, from their review of both experimental studies and the impact of high stakes tests in naturalistic studies, that those who are motivated by external examinations are likely to have performance goals and not learning goals. Students with performance goals are 'shallow' learners who make a great deal of use of rote learning as compared with those with learning goals.

Crooks (1988) and **Black and Wiliam** (1998) were extensive reviews of classroom assessment practices. Crooks looked at the impact of assessment on students, including self-efficacy, intrinsic motivation and attribution of success or failure. He emphasised the need to pay attention to motivational outcomes. His findings in the areas common with this review are entirely consistent with conclusions from the studies reviewed here. Black and Wiliam identified the features of classroom assessment which were found to be associated with raising levels of achievement. Some of these closely parallel the actions reported here as tending to decrease the negative and increase the positive impact of summative assessment, in particular the teacher providing non-ego-involving feedback and involving students in self-assessment.

The review by **Natriello** (1987) covered a wider range of assessment purposes than the Crooks and Black and Wiliam reviews. However, since this review pre-dated the strong influence of high stakes testing it does not reflect the later research on its impact. However, Natriello pointed out that, although clarifying goals for learning for students is important, too much emphasis in specific assessment criteria can encourage students to focus only on what is assessed. Although Ames' review was not concerned with the impact of high-stakes testing, her main conclusions about classroom assessment support the findings of this review in relation to the impact of assessment with a summative purpose on students' motivation.

The review by **McDonald** (2001) was specifically focused on test anxiety and its impact on students' performance. As in the current review, females were found to score more highly on test anxiety that males. In relation to performance, there was considerable evidence from a range of countries and across academic subjects of a negative relation between test anxiety and test

performance. Although there were also studies which reported no relationship, McDonald concludes that overall the influence is negative and large enough to make the difference between passing and failing a test for at least one fifth of the students. Thus this review supports the importance of paying attention to test anxiety.

The reviews of **Madaus and Clarke** (1998) and **McNeil and Valenzuela** (2000) provide more information about minority students than was found in the studies considered in this review, where only one paper specifically mentioned minority students. In the paper by Roderick and Engel (2001) the 'at risk' students were all Afro-American or Latino. The extent to which the tests motivated some of these students to make more effort is uncertain, since the consequences of failure (being held back in their grade and not promoted to the next grade with their peers) were very severe for the students. Moreover, those who improved their performance on the tests were those who had considerable support from helpful adults.

McNeil and Valenzuela (2000) note how the Texas Assessment of Academic Skills promotes one mode of teaching and learning, that of mastering discreet, often disconnected, pieces of information. In this way of teaching, students' cultures and links to everyday life are ignored. In essence, the McNeil and Valenzuela review reports extreme versions of the tendencies that have emerged from the present review, whilst emphasising the far greater negative impact on students who are already disadvantaged.

7.4 Meanings of the review for different user groups

Summaries of the review for different user groups have been prepared by teachers, headteachers and assessment advisers. A summary providing students' perspectives on the review findings has also been prepared.

7.5 Unanswered questions

Some questions which are raised by the review but not answered by the studies are:

- What is the cause-effect relationship between aspects of motivation and the impact of summative assessment?
- Is the impact of testing on motivation reversible or irreversible? Does this vary for students of different age and level of achievement?
- Does summative assessment *per se* have an impact on motivation or is any impact a consequence of the results having high stakes?
- What examples are there of practice where summative assessment serves a positive purpose without negative impact on motivation?
- How can summative assessment be used to report on students' achievement in a way that avoids some students inevitably experiencing a demotivating sense of failure?

8. CONCLUSIONS AND RECOMMENDATIONS

8.1 Conclusions and recommendations for assessment practice and policy

Several of the points in Chapter 6 indicate steps that can be taken to move practice in summative assessment in directions that avoid the detrimental impact on students' motivations for learning. In developing these findings into specific action, the authors have drawn upon the findings of the 19 studies, other writing in commentaries and reviews of research relating to assessment which informed the background to the review, and the outcomes of the consultation conference held with policy-makers and practitioners from all parts of the UK.

There are clear messages for how the negative impact of summative assessment on motivation for learning can be minimised. In some cases these refer to practices that should be ended as far as possible and, in particular, to the following:

- To avoid drill and practice for tests
- To de-emphasise tests by using a range of forms of classroom assessment and recognising the limitations of tests
- To prevent the content and methods of teaching from being limited by the form and content of tests
- To avoid children being faced with tests in which they are unlikely to succeed

However, rather than indicate only what should be avoided, the review has identified more positive messages that identify action that can be taken to ensure that the benefits of summative assessment can be had without associated impact on students' motivation for learning. These recommendations derive from both the research studies, and from discussion of the review findings and reference to current practice in the UK at the consultation conference.

8.1.1 Practice

(a) Reduce the narrowing impact on the curriculum and on teaching methods by professional development that emphasises learning goals and learner-centred teaching approaches.

(b) Share and emphasise learning goals, not performance goals, with students and provide feedback to students in relation to these goals.

(c) Share in developing and implementing a school-wide policy that includes assessment both *for* learning (formative) and *of* learning (summative) and ensure that the purpose of all assessment is clear to all involved, including parents and students.

(d) Develop students' understanding of the goals of their learning, the criteria by which is it assessed and their ability to assess their own work.

(e) Implement strategies for encouraging self-regulation in learning and positive interpersonal relationships. Ways of doing this have been developed through research by McCombs, 1999. These include strategies for:

- fostering caring personal relationships among students and between students and teacher
- helping students to challenge themselves and think for themselves whilst learning
- helping students to take some responsibility for learning to learn
- supporting students in directing their own learning
- providing students with some choice and control over their learning process
- encouraging collaboration and the use of other students as resources for learning

(f) Avoid comparisons between students based on test results.

(g) Present assessment realistically, as a process which is inherently imprecise and reflexive, with results that have to be regarded as tentative and indicative rather than definitive.

8.1.2 Policy

(a) Recognise that current high stakes testing is failing to provide valid information about students' attainment for three main reasons:

- Tests are too narrowly focused to provide information about students attainment.
- This would be the case even if the tests were reliable. There is, however, evidence on unreliability on two counts:
  (i) any assessment or test is subject to large errors of measurement which are often not revealed;
  (ii) the consequence of teaching to the test means that students may not in reality have the skills or understanding which the test is designed to assess, since teachers are driven by the high stakes to teach students how to pass tests even when they do not have this skills and understanding.
- The reliability of the tests as useful indicators of student attainment is also being undermined by the differential impact of the testing procedures, and particularly repeated practice tests on a significant proportion of students. Thus girls and low achieving students are likely to have high level of test anxiety which can influence their performance. More importantly, however, is that older and lower achieving students are likely to minimise effort and may even answer randomly since their experience is that they are failing anyway. Thus results are unreliable and exaggerate the difference between the high and low achievers.

(b) Recognise the importance for students' attainments of outcomes of education that relate to the components of motivation. Not only is there a growing recognition of the value of learning to learn and of the drive and energy to continue learning, but there is empirical evidence that these are

positively related to attainment. For example, the OECD/PISA (2001) provides firm evidence that achievement of literacy is positively related to students' interest in their learning, the extent to which their learning strategies help them to develop understanding through linking to existing knowledge instead of just memorising, and the extent to which they feel in control of their learning.

(c) Provide professional development, particularly for senior school management, aimed at enabling schools to develop a range of assessment strategies and using summative information of different kinds for improving the learning of their students. Current training focuses too narrowly on the use of test scores and accountability and target setting and needs to be more learner-focused.

(d) For summative purposes in reporting on individual students, move towards testing students when their teachers judge them to be ready to show their achievement at a certain level, as in the Scottish system for national testing in the 5–14 programme, thus avoiding experience of failure and its impact on self-esteem.

(e) Ensure that the criteria used in school evaluation, including self-evaluation, make explicit reference to a full range of subjects and including moral, spiritual and cultural as well as cognitive aims and an appropriate variety of teaching methods and learning outcomes.

(f) Develop schools' self-evaluation practices including teachers' assessment skills through targeted professional development.

(g) For tracking national standards, sample students rather than testing all and use a wider range of test forms and items.

(h) Quantify the 'cost' of current practice, including teaching time taken up for testing and practice testing and adding to teachers' workloads by additional marking in addition to the cost of the tests and their development.

(i) Use test development expertise to create new tests and assessment that will enable all valued outcomes of education, including creativity and learning to learn to be assessed.

(j) Reduce the 'stakes' of summative assessment by avoiding comparisons among schools in terms of test results; and end the practice of basing targets only on test results.

8.2 Conclusions and recommendations for assessment research

The finding that only 19 studies dealing with the impact of summative assessment on motivation for learning emerged from the search carried out indicates that this is an under-researched area. A large corpus of research on cognitive outcomes of educational practice and indeed of assessment, evaluation and testing, exists. The number of research studies concerned with affective and conative (ie mental activity) outcomes of assessment is very small by comparison. Yet these outcomes are recognised as being of growing

significance. One set of reasons for this is that these outcomes relate to the development of attitudes and the release of energy that promotes continued learning throughout life, as needed by citizens of a world in which the pace of change is not just continuing but is accelerating. Other reasons are more empirically based. Research indicates that the development of broad understanding and applicable skills follows when students learn in certain ways and use their energy for developing understanding rather than memorising. For example, in the findings of the OECD/PISA study (2001), the achievement of literacy has been found to be positively related to students' interest in what they are learning, to the extent to which their learning strategies help them to create links between new and existing knowledge and to the extent to which they feel in control of their learning. All these are feature of learning which the research reviewed here has shown can be endangered by certain aspects of summative assessment.

Thus there is a need to extend the research base that informs policy and practice in these matters. Calls for further research to clarify or verify findings of the studies reviewed has in some cases been made by the authors themselves. Thus Perry (1998) calls for the development of more valid measures of young children's motivation and of measures that enable qualitative comparisons to be made of children's writing. Leonard and Davey (2001) indicate that more research should take into account the views and viewpoints of students. Brookhart and DeVoge (1999) acknowledge that their research was carried out in classrooms that were positive assessment environments and need to be repeated in a wider range of classrooms.

The reviewers identified the need for research to fill gaps in existing research and to address additional questions arising from several studies. Gaps arise due to the complexity of the motivational process (see Figure 3) and the need for a theory of motivation which relates to everyday life as well as to learning in schools. Until the area is better underpinned by theory, it is unlikely for measures to be developed which address its complexity.

However in addition to better theory-based research tools, the impact of assessment on motivation requires studies of types that enable the relationship between the independent variable of assessment practice and the dependent variable of motivation to be investigated. The research deals either with the natural context and a naturalistic intervention, where variables cannot be controlled, or sets up experimental situations, where variable can be controlled or randomised but where the relevance to real situations may be unclear. There is room for improved studies of both these kinds. In the 'real' situation, research is needed that

- goes beyond associations between variables and enables the direction of cause to be explored
- focuses on the direct impact of assessment rather than depending on surrogates or assumptions that certain circumstances have been brought about by assessment of testing practice
- involves longitudinal or cohort studies that are capable of disentangling the effect of maturity from that of the independent variable. (In practice, this is extremely difficult in the context of a naturalistic intervention which affects all students and so prevent the use of control groups. Such situations call for complex designs in which, for instance, there is variation in the nature of the intervention).

- compares cognitive, affective and conative outcomes in students in countries where high stakes testing is prevalent and those (such as Norway) where there is no high stakes national testing

Whilst some of these problems could be answered theoretically by well-designed experimental studies, it is difficult to reconcile the need for relevance to the complexity of classroom work with the need to set up controlled conditions that enable cause and effect to be investigated. Moreover, it would seem unethical to set up as trials some of the practices that have been found to occur in real classrooms and which negatively impact on students' motivation for learning.

9. REFERENCES

9.1 Included studies

Benmansour N (1999) Motivational orientations, self-efficacy, anxiety and strategy use in learning high school mathematics in Morocco. *Mediterranean Journal of Educational Studies*. 4: 1-15

Brookhart S, DeVoge J (1999) Testing a theory about the role of classroom assessment in student movitation and achievement. *Applied Measurement in Education* 12: 409-425

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

Davies J, Brember I (1999) Reading and mathematics attainments and self-esteem in years 2 and 6: an eight year cross-sectional study. *Educational Studies* 25: 145-157

Davies J, Brember I (1998) National curriculum testing and self-esteem in year 2 the first five years: a cross-sectional study. *Educational Psychology* 18: 365-375

Duckworth K, Fielding G, Shaughnessy J (1986) *The relationship of high school teachers' class testing practices to students' feelings of efficacy and efforts to study.* US: Oregon University

Evans E, Engelberg R (1988) Students perceptions of school grading, *Journal of Research and Development in Education* 21: 44-54

Ferguson C, Francis J (1979) Motivation and mode: an attempt to measure the attitudes of 'O' level GCE candidates to English language. *Educational Studies* 5: 231-239

Gordon S, Reese M (1997) High stakes testing: worth the price? *Journal of School Leadership* 7: 345-368

Hughes B, Sullivan H, Beaird J (1986) Continuing motivation of boys and girls under differing evaluation conditions and achievement levels. *American Educational Research Journal* 23: 660-667

Johnston J, McClune W (2000) *Selection project sel 5.1: Pupil motivation and attitudes - self-esteem, locus of control, learning disposition and the impact of selection on teaching and learning.* Belfast: Queen's University

Leonard M, Davey C (2001) *Thoughts on the 11 plus.* Belfast: Save the Children Fund

Little A (1994) Types of assessment and interest in learning: variation in the south of England in the 1980s. *Assessment in Education* 1: 201-222

Paris S, Lawton T, Turner J, Roth J (1991) A developmental perspective on standardised achievement testing. *Educational Researcher* 20: 12-20

Perry N (1998) Young children's self-regulated learning and contexts that support it. *Journal of Educational Psychology* 90: 715-729

Pollard A, Triggs P, Broadfoot P, Mcness E, Osborn M (2000) *What pupils say: changing policy and practice in primary education* (chapters 7 and 10). London: Continuum

Reay D, Wiliam D (1999) 'i'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal* 25: 343-354

Roderick M, Engel M (2001) The grasshopper and the ant: motivational responses of low achieving pupils to high stakes testing. *Educational Evaluation and Policy Analysis* 23: 197-228

Schunk D (1996) Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal* 33: 359-382

9.2 Excluded studies

Alsaker F (1989) School achievement, perceived academic competence and global self-esteem. *School Psychology International* 10: 147-158

Ames C (1992) Classrooms: goals, structures and student motivation. *Journal of Educational Psychology* 84: 261-271

Ames C (1990) Motivation: what teachers need to know. *Teachers College Record* 91: 409-421

Anderson L (1997) Educational testing and assessment: lessons from the past, directions for the future. *International Journal of Educational Research* 27: 355-445

Assessment Reform Group (1999) *Assessment for learning: beyond the black box.* Cambridge: University of Cambridge, School of Education

Bagley C, Mallick K (1996) Cross-cultural studies of self-esteem, self-differentiation and stress in school students. *Research in Education* 56: 21-30

Battle J (1997) The relative effects of married versus divorced family configuration and socioeconomic status on the educational achievement of African American middle-grade students. *Journal of Negro Education* 66(1): 29-42

Becker W, Rosen S (1992) The learning effect of assessment and evaluation in high school. *Economics of Education Review* 11(2): 107-118

Benware C, Deci E (1984) Quality of learning with an active versus passive motivational set. *American Educational Research Journal* 21: 755-765

Beron (1990) Joint determination of current classroom performance and additional economics classes: a binary/continuous model. *Journal of Economic Eduction* 21(3): 255-264

Bishop J (1989) *Incentives for learning: Why American high school students compare so poorly to their counterparts overseas.* State University of New York, Ithaca. School of Industrial and Labour Relations at Cornell University, US: working paper 89-09

Black P (2000) Research and the development of educational assessment. *Oxford Review of Education* 26: 407-419

Black P, Wiliam D (1998) Assessment and classroom learning. *Assessment in Education* 5: 7-74

Black P, Wiliam D (1998) *Inside the black box.* London: King's College London, School of Education

Boaler J, Wiliam D, Brown M (2000) Students' experiences of ability grouping: disaffection, polarisation and the construction of failure. *British Educational Research Journal* 26: 631-648

Broadfoot P, Pollard A, Osorn M, McNess E, Triggs P (1998) Categories, standards and instrumentalism: theorising the changing discourse of assessment policy in English primary education. Paper presented at the Annual meeting of the American Educational Research Association, April 13-17. San Diego, California, US

Broadfoot P (1979) Communication in the classroom: a study of the role of assessment in motivation. *Educational Review* 31: 3-10

Broadfoot P (1998) Records of achievement and the learning society: a tale of two discourses, *Assessment in Education* 5: 447-477

Bronzaft A, Murgatroyd D, McNeilly R (1974) Test anxiety among black college students: a cross-cultural study. *Journal of Negro Education* 43: 190-193

Brown M (1989) Graded assessment and learning hierarchies in mathematics: an alternative view. *British Educational Research Journal* 15: 121-128

Brown A, Campione J, Webber L, McGilly K (1992) In Gifford BOCM (ed) *Changing assessments alternative views of aptitude, achievement and* instruction. London: Kluwer Academic Publishers

Brown S, Walberg H (1993) Motivational effects on test scores of elementary students.*Journal of Educational Research* 86: 133-136

Bryant D (1996) *A comparison of multiple-choice versus alternative assessments: strengths and limitations.* New York, US: University of the State of New York

Cain K, Dweck C (1995) The relation between motivational patterns and achievement cognitions through the elementary years. *Merrill-Palmer Quarterly* 41: 25-52

Cameron J, Pierce W (1994) Reinforcement, reward and intrinsic motivation: a meta- analysis. *Review of Educational Research* 64: 363-423

Cavallo A (1992) The retention of meaningful understanding of meiosis and genetics. Paper presented at a poster session at the Annual Conference of the National Association for Research in Science Teaching, March 22. Boston, Massachusetts, US

Chongqin L (2001) The experimental research on the fostering of students' learning achievement motivation. *Pyschological Science* 24: 377

Cooper B (1996) Using data from clinical interviews to explore students' understanding of mathematics test items: relating Bernstein and Bourdieu on culture to questions of fairness in testing. Paper presented at the American Educational Research Association meeting, New York, US

Covington M, Muelluer K (2001) Intrinsic versus extrinsic motivation: an approach/avoidance reformulation. *Educational Psychology Review* 13: 157-176

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

Daniels D, Kalkman D, McCombs B. (2001) Young children's perspectives on learning and teacher practices in different classroom contexts: implications for motivation. *Early Education and Development* 12: 253-273

Dart B, Burnett P, Cambell-Purdie N, Boulton-Lewis G, Campbell J, Smith D (2000) Students' conceptions of learning, the classroom environment and approaches to learning. *The Journal of Educational Research* 93: 262-270

Dart B, Burnett P, Boulton-Lewis G, Cambell J, Smith D, McCrindle A (1999) Classroom learning environments and students' approaches to learning. *Learning Environments Research* 2: 137-156

Daugherty R, Freedman ES (1998) Tests, targets and tables: the use of key stage 2 assessment data in Wales. *The Welsh Journal of Education* 7: 5-21

Davies J, Brember I (1999b) Reading and mathematics attainments and self-esteem in years 2 and 6: an eight year cross-sectional study. *Educational Studies.* 25: 145-157

Deci E, Ryan R (1985) *Intrinsic motivation and self determination in human behaviour.* New York, US: Plenum

Denscombe M (2000) Social conditions for stress: young people's experience of doing GCSEs. *British Educational Research Journal* 26: 359-374

Dixon D (1990) Organizational culture and correlates of effectiveness in California's dropout prevention programs, Annual Meeting of the American Educational Research Association, California, US

Eaton J, Dembo M (1997) Differences in the motivational beliefs of Asian American and non-Asian students. *Journal of Educational Psychology* 89: 433-440

Eccles J, Wigfield A, Midgley C, Reuman D, MacIver D, Feldlauger H (1993) Negative effects of traditional middle schools on students' motivation. *The Elementary School Journal* 93: 553-574

Elliot A, Covington M (2001) Approach and avoidance motivation. *Educational Psychology Review* 13: 73-92

Elliot E, Dweck C (1988) Goals: an approach to motivation and achievement. *Journal of Personality and Social Pschology*

Elliott J, Hufton N, Hildreth G, Illushin L (1999) Factors influencing educational motivation: a study of attitudes, expectation and behaviour of children in Sunderland, Kentucky and St Petersburg. *British Educational Research Journal* 25: 75-94

Entwistle N, Kozeki B (1985) Relationship between school motivation, approaches to studying, and attainment among British and Hungarian adolescents. *British Journal of Educational Psychology* 55: 124-137

Ferriman B, Lock R, Soares A (1993) Testing science at Ks3. *Forum* 35: 79-81

Firestone W, Mayrowetz D, Fairman J (1998) Performance-based assessment and intructional change: the effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis* 20: 95-113

Firestone W, Mayrowetz D (2000) Rethinking 'high stakes': lessons from the United States and England and Wales. *Teachers College Record* 102: 724-749

Fortier M, Vallerand R, Guay F (1995) Academic motivation and school performance: towards a structural model. *Contemporary Educational Psychology* 20: 257-274

Gad Y (2000) Reforming motivation: how the structure of instruction affects students' learning experiences. *British Educational Research Journal* 26: 191-210

Ganguly I (1993) The motivational effect of televised instruction on teacher directed science learning. Annual Conference of the International Visual Literacy Association, Ohio, US

Goins B (1993) Student motivation.*Childhood Education* 69(5): 316-317

Gose B (2000) *More points for 'strivers': the new affirmative action?* New York, US: ERIC Report

Gottfried A, Fleming J, Gottfried A (1994) Role of parental motivational practices in children's academic intrinsic motivation and achievement. *Journal of Educational Psychology* 86: 104-113

Goudas M, Dermitzaki I, Bagiatis K (2000) Predictors of students' intrinsic motivation in school physcial education. *European Journal of Psychology of Education* 15: 271-280

Goudas M, Biddle S (1994) Perceived motivational climate and intrinsic motivation in school physical education classes. *European Journal of Psychology of Education* 9: 242-248

Green L (1990) Test anxiety, mathematics anxiety and teacher comments: relationships to achievement in mathematics classes. *Journal of Negro Education* 59: 320-335

Green R, Griffore R (1980) The impact of standardized testing on minority students. *Journal of Negro Education* 49: 238-252

Greeno J, Pearson P, Schoenfield A (1996) *Implication for NAEP of research on learning and cognition.* Stnaford, California, US: Institute for Research on Learning

Hallinan P, Danaher P (1994) The effect of contracted grades on self-efficacy and motivation in teacher education courses. *Educational Research* 36: 75-82

Hancock D (2001) Effects of test anxiety and evaluative threat on students' achievement and motivation. *The Journal of Educational Research* 94: 284-290

Hembree R (1988) Correlates, causes, effects and treatment of test anxiety. *Review of Educational Research* 58: 47-77

Ho H, Senturk D, Lam A, Zimmer J, Hong S, Okamoto Y (2000) The affective and cognitive dimensions of math anxiety: a cross-national study. *Journal for Research in Mathematics Education* 31: 362-379

Houtz L (1995) Instructional strategy change and the attitude and achievement of 7th grade and 8th grade science students. *Journal for Research in Science Teaching* 32: 629-648

Hughes K (1989) *The children's academic motivation inventory: Validation evidence for generalization to a high school population*. Alabama, US

Institute for Academic Excellence (1997) *Toward a balanced approach to reading motivation: resolving the intrinsic-extrinsic rewards debate. Report.* Madison, Wisconsin, US: The Institute for Academic Excellence

Ireson J (1999) Ability grouping in the secondary school: the effects on academic achievement and pupils' self-esteem. Paper presented at the British Educational Research Association Annual Conference, September 2-5: University of Sussex

Ireson J, Hallam S, Plewis P (2001) Ability grouping in secondary schools: effects on pupils' self-concepts. *British Journal of Educational Psychology* 71: 315-326

James C, Tanner C (1993) Standardised testing of young children. *Journal of Research and Development in Education* 26: 143-152

James M (2000) Measured lives: the rise of assessment as the engine of change in English schools. *The Curriculum Journal* 11: 343-364

James M, Gipps C (1998) Broadening the basis of assessment to prevent the narrowing of learning. *The Curriculum Journal* 9: 285-297

Johnston P, Guice S, Baker K, Malone J, Michelson N (1995) Assessment and teaching and learning in 'literature-based' classrooms. *Teaching and Teacher Education* 11: 359-371

Keith T, Cool V (1988) *Testing theories of learning: Effects on high school achievement.* Virginia, US

Kellaghan T (2001) The future and challenges of educational assessment in the 21st century. International Conference of the International Association for Educational Assessment, Rio de Janerio, Brazil

Kellaghan T, Madaus G, Raczek A (1996) *The use of external examinations to improve student motivation.* Washington DC, US: AERA

Keltikangas J (1992) Self esteem as a predictor of future school achievement. *European Journal of Psychology of Education* 7: 123-130

Kingdon M (2001) The future and challenges of educational assessment in the 21st century. International Conference of the International Association for Educational Assessment. Rio de Janerio, Brazil

Kiplinger V, Linn R (1996) Raising the stakes of test administration: the impact on student performance on the national assessment of educational progress. *Educational Assessment* 3: 111-133

Kohn A (1994) *The risks of rewards.* Illinois, US: ERIC Clearinghouse of Elementary and Early Childhood Education

Kohn A (1996a) Grading: the issue is not how but why. *Educational Leadership* 52(2): 38-41

Kohn A. (1996b) By all available means: Cameron and Pierce's defence of extrinisic motivators. *Review of Educational Research* 66: 1-4

Kohn A (1999a) The costs of overemphasizing achievement. *The School Administrator*: 40-46

Kohn A (1999b) From degrading to de-grading. *High School Magazine* 6(5): 38-43

Kohn A (2000) *The case against standardized testing*. Portsmouth, NH: Heinemann

Koretz D, Barron S, Keith S (1966) *Final report: perceived effects of the maryland school performance assessment program.* Los Angeles, US: National Center for Research on Evaluation Standards and Student Testing (CRESST)

Krapp A (1999) Intrinsic learning motivation and interest research approaches and conceptual considerations. *Zeitschreift fur Padagogik* 45: 387-406

Lange G, Adler F (1997) Motivation and achievement in elementary children. Paper presented at the Biennial Meeting of the Society for Research in Child Development, April 3-6: North Carolina, US

Lepper M, Keavney M, Drake M (1996) Intrinsic motivation and extrinsic rewards: a commentary on Cameron and Pierce's meta-analysis. *Review of Educational Research* 66: 5-32

Liberatore C, Schafer L (1994) Relationships between students' meaningful learning orientation and their understanding of genetics topics*. Journal of Research in Science Teaching* 31(4): 393-418

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

Madaus G (1991) The effects of important tests on students: implications for a national examination system. *Phi Delta Kappan* November 1991: 226-231

Madaus G (1993) A national testing system: manna from above? An historical/technological perspective. *Educational Assessment* 1**:** 9-26

Madaus G, Clarke M (1988) The influence of testing on the curriculum. In Tanner (ed) *Critical issues in curriculum, 87th year book of NSSE part 1* University of Chicago Press, Chicago, US: 83-121

Madaus G, Clarke M (1998) The adverse impact of high stakes testing on minority students: evidence from 100 years of test data. High Stakes K-12 Testing Conference, Harvard University, US

Maehr M, Anderman E (1993) Reinventing schools for early adolescents: emphasising task goals. *The Elementary School Journal* 93: 593-609

Marzano R (1998) *A theory based meta analysis of research on instruction.* Colorado, US: Office of Educational Research and Improvement

Mathews J (1998) *Class struggle. What's wrong (and right) with America's best public high schools?* New York, US: Times Books

McCombs B, Lauer P (1997) Development and validation of the learner-centred battery: Self assessment tools for teacher reflection and professional development. *The Professional Educator* 20: 1-20

McDonald A (2001) The prevalence and effects of test anxiety in school children. *Educational Psychology* 21: 89-101

McInerny D, Rocher L, McInerny V, Marsh H (1997) Cultural perspectives on school motivation: the relevance and application of goal theory. *American Educational Resarch Journal* 34: 207-236

McMenniman M (1989) In Langford P (ed), *Educational psychology: an Australian* perspective. Cheshire: Longman

McNeil L, Valenzuela A (2000) *The harmful impact of the TAAS system of testing in Texas: beneath the accountability rhetoric.* Houston, Texas, US: Rice University

Middleton J (1992) Teachers' vs. students' beliefs regarding intrinsic motivation in the mathematics classroom: a personal constructs approach. Paper presented at the Annual Meeting of the American Educational Research Association, April 22. San Francisco, California, US

Miller D, Seraphine A (1993) Can test scores remain authentic when teaching to the test? *Educational Assessment* 1: 119-129

Mitchell M (1992) Situational interest: its multifaceted structure in the secondary mathematics classroom. Annual Meeting of the American Educational Research Association, California, US

Moely B (1995) Cross-sectional and longitudinal assessments of changes in motivational beliefs of elementary and middle school children. Annual Meeting of the American Educational Research Association, Louisiana, US

Moshe N, Hagit L, McKeachie W, Yi-Guang L (1997) Individual differences in students' retention of knowledge and conceptual structures learned in university and high school courses: the case of test anxiety. *Applied Cognitive Psychology* 11: 507-526

Murphy P (1994) Assessment: gender implications. Paper presented at a conference in Dublin. Dublin: University of Dublin Teaching Committee

Natriello G (1987) The impact of evaluation processes on students. *Educational Psychologist* 22: 155-175

Nenniger P (1999) On the role of motivation in self directed learning: the 'two-shells-model of motivated self-directed learning' as a structural explanatory concept. *European Journal of Psychology of Education* 14: 71-86

Nichol J (1984) Achievement motivation, conceptions of ability subjective experience, task, choice and performance. *Pyschological Review* 91: 328-346

Norwich B (1999) Pupils' reasons for learning and behaving and for not learning and behaving in English and maths lessons in a secondary school. *British Journal of Educational Psychology* 69: 547-569

Noss R, Goldstein H, Hoyles G (1989) Graded assessment and learning hierarchies in mathematics. *British Educational Research Journal* 15: 109-120

Okun M, Fournet L (1993) Academic self-esteem and perceived validity of grades: a test of self-verification theory. *Contemporary Educational Psychology* 18: 414-426

O'Neil H, Sugrue B, Baker E (1998) *Effects of motivational interventions on the national assessment of educational progress mathematics performance.* Washington, US: Office of Educational Research and Improvement

Pajares F, Miller D (1998) Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *Journal of Experimental Education* 65(3): 213-228

Parkes J (1997) Performance assessment and student motivation: questioning construct relevant variance. Annual Meeting of the American Educational Research Association, Pennsylvania, US

Paulsen M, Feldman K (1999) Student motivation and epistemological beliefs. *New Directions for Teaching and Learning* 78: 17-26

Perrin B (1998) Effective use and misuse of performance measurement. *The American Journal of Evaluation* 19: 367-379

Perrin M (1989) Switzerland: Summative evaluation and pupil motivation. In Weston P (ed) *Assessment of pupil achievement: motivation and school success.* Report of the Educational Research workshop, September 2-15, Liege, Belgium. Amsterdam: Swets and Zeitlinger

Pintrich P (1999) The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research* 31: 459-470

Pitman J (2001) A new certification deal for 15-19 year olds. The International Association for Educational Assessment, Rio de Janerio, Brazil

Pitman J, Dudley P (1985) Criteria based assessment: the Queensland experience. International Conference of the International Association for Educational Assessment, Oxford

Randhawa B (1993) Role of mathematics self-efficacy in the structural model of mathematics achievement. *Journal of Educational Psychology* 85(1): 41-48

Raven J (1991) In Weston P (ed) *Assessment of pupil achievement: motivation and success*: 45-78. Amsterdam: Swets and Zeitlinger, Lisse

Reap M, Cavallo A (1992) Students' meaningful understanding of science concepts: gender differences. Annual Conference of the National Association for Research in Science Teaching, Oklahoma, US

Reuman D, MacIver D (1994) *Effects of instructional grouping on seventh graders' academic motivation and achievement*. Maryland, US: Centre for Research on Effective Schooling for Disadvantaged Students

Rheinberg F, Vollmeyer R, Burns B (2001) QCM: a questionnaire to assess current motivation in learning situations. *Diagnostica - Gottingen-issn* 47: 57-66

Roth J, Brooks-Gunn J, Murray L, Foster W (1998) Promoting healthy adolescents: synthesis of youth development program evaluations. *Journal of Research on Adolescents* 8(4): 423-459

Ryan R, Deci E (2000) Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology* 25: 54-67

Salmi H (1993) *Science centre education. Motivation and learning in informal education*. Finland: Academic Dissertation, University of Helsinki

Saul M (1997) Common sense: the most important standard. *Mathematics Teacher* 90(3): 182-184

Schunk D (1991) Self-efficacy and academic motivation. *Educational Psychologist* 26: 207-231

Schunk D, Swartz C (1993) Goals and progress feedback: effects on self-efficacy and writing achievement. *Contemporary Educational Psychology* 18: 337-354

Sheets R (1995) *College board advanced placement Spanish literature for at-risk native speakers: a model with multicultural, bilingual and gifted dimensions.* Washington, US

Sheldon K, Bruce J (1998) Standards, accountability and school reform: perils and pitfalls. *Teachers College Record* 100: 164-180

Shepard L (1991) Will national tests improve student learning? *Phi Delta Kappan,*: 232-238

Shepard L (2000) The role of assessment in a learning culture. American Educational Research Association, New Orleans, US

Skaalvik E (1994) Attribution of perceived achievement in school in general and in maths and verbal areas: relations with academic self-concept and self-esteem. *British Journal of Educational Psychology* 64: 133-143

Smith I (1993) *An investigation into students' perceptions of the learning environment provided by hypermedia tools in an interdisciplinary high school course of studies.* Oregon, USA

Smith M (1991) Put to the test: the effects of external tests on teachers. *Educational Researcher* 20: 8-11

Smith T (1991) Agreement of adolescent educational expectations with perceived maternal and paternal educational goals. *Youth & Society* 23(2): 155-174

Snow R, Jackson D (1992) *Assessment of cognitive educational processes and outcomes: status report of empirical studies. Project 2.3: Enhancing the utility of performance assessments.* California, US: National Center for Research on Evaluation Standards and Student Diversity

Stiggins R (1992) Two disciplines of educational assessment. Bolder, Colorado: Education Commission of the States Assessment Conference

Stiggins R (1999) Assessment, student confidence and school success. *Phi Delta Kappan* 81(3): 191-198

Stobart G (2001) The validity of national curriculum assessment. *British Journal of Educational Studies* 49: 26-39

Stone CA, Lane S (1999) MSPAP performance gains from 1993-98 and their relationship to 'MSPAP impact' and school characteristic variables. Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 19-23. Montreal, Canada

Stubblebine P (1998) Effect of threatening feedback on expected grade, self-efficacy, and motivation. *Perceptual and Motor Skills* 86: 67-77

Theophilides C, Dionysiou O (1996) The major functions of the open book examination at the university level: a factor analytic study. *Studies in Educational Evaluation* 22: 157-170

Thomas S, Madaus G, Raczek A, Smees R (1998) Comparing teacher assessment and standards task results in England: the relationship between pupil characteristics and attainment. *Assessment in Education* 5: 213-240

Torrance H, Pryor J (1995) Investigating teacher assessment in infant classrooms: methodological problems and emerging issues. *Assessment in Education* 2: 305-320

Vredeveld G, Jeong J (1990) Market efficiency and student-teacher goal agreement in the high school economics course: a simultaneous choice modeling approach. *Journal of Economic Education* 21(3): 317-335

Watkins C, *et al.* (2001) Learning about learning improves performance. *National School Improvement Network Research Matters* 13

Weinberger E, McCombs B (2001) The impact of learner centred practices on the academic and non academic outcomes of upper elementary and middle school students. Seattle, US: American Educational Research Association

Wiggins A, Tymms PB (2000) Dysfunctional effects of public performance indicator systems: a comparison between English and Scottish primary schools. Paper presented at the European Conference on Educational Research, 20-23 September. Edinburgh

Williams J (1996) The relation between efficacy for self-regulated learning and domain-specific academic performance, controlling for test anxiety. *Journal of Research and Development in Education* 29: 77-81

Williams J, Ryan J (1999) National testing and the improvement of classroom teaching: Can they coexist? *British Journal of Educational Research* 26: 49-74

Williams M (1993) *Interactions among attributional style, attributional feedback and learner-controlled CBJ.* Minnesota, US

Williams R, Dotson W, Don P, Williams W (1980) The war against testing: a current status report critical issues in testing and achievement of black Americans. *Journal of Negro Education* 49: 263-273

Wilson M (1992) Educational leverage from a political necessity: implications of new perspectives on student assessment for chapter 1 evaluation. *Educational Evaluation and Policy analysis* 14: 123-144

Wolters C, Pintrich P (1998) Contextual differences in student motivation and self-regulated learning in mathematics, English and social studies classrooms. *Instructional Science* 26: 27-47

Wolters A, Yu S, Pintrich P (1996) The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences* 8: 211-237

Worrall N (2001) Test, testing, testing: investigating student attitudes towards, and perceptions of, eleven years of testing and target setting. *Forum* 43: 13-18

Yair G (2000) Reforming motivation: how the structure of instruction affects students' learning experiences. *British Educational Research Journal* 26: 191-210

Zimmerman B, Bandura A, Martinez-Pons M (1992) Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting, *American Educational Research Journal,* 29, 663-676

9.3 Other references used in the text

Ames C (1990) Developing a learning orientation. Paper presented at annual meeting of the AERA, Boston, US

Clarke M, Madaus GF, Horn CJ, Ramos MA (2000) Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies* 32(2) 159-181

Deci EL, Koestner R, Ryan RM (1999) A meta-analysis review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125: 627-688
DfEE (2000) *The National Curriculum,* London: DfEE and QCA

Dweck CS (1992) The study of goals in psychology. *Psychological Science* 3: 165-167

EPPI-Centre (2001) *Guidelines for extracting data and quality assessing primary studies in educational research (version 0.94).* London: EPPI-Centre, Social Science Research Unit

Grolnick W.S, Ryan RM (1987) Autonomy in children's learning: an experimental and individual difference investigation. *Journal of Personality and Social Psychology* 52: 890-898

Harlen W, James M (1997) Assessment and learning: difference and relationships between formative and summative assessment. *Assessment in Education* 4(3) 365-379

Hidi S (2000) An interest researcher's perspective: the effects of extrinsic and intrinsic factors on motivation. In Sansome C, Harackiewicz JM (eds) *Intrinsic and extrinsic motivation: the search for optimal motivation and performance.* New York, US: Academic Press

Hidi S, Harackiewicz JM (2000) Motivating the academically unmotivated: a critical issue for the 21st century. *Review of Educational Research* 70(2): 151-179

Johnston C (1996) *Unlocking the will to learn.* Thousand Oaks, California, US: Corwin Press

Katzell RA, Thompson DE (1990) Work motivation: theory and practice. *American Psychologist* 45: 144-153

Kohn A (1993) *Punished by rewards.* Boston, US: Houghton Mifflin

Koretz D (1988) Arriving at Lake Wobegon: are standardised tests exaggerating achievement and distorting instruction? *American Educator* 12 (2): 8-15, 46-52

Koretz D, Linn L, Dunbar SB, Shepard LA (1991) The effects of high-stakes testing on achievement: preliminary findings about generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, US

Linn R, Dunbar S, Harnisch D, Hastings C (eds) (1982) *The validity of the Title 1 evaluation and reporting systems.*California: Beverley Hills, US: Sage Publications

Little A (1994) Types of assessment and interest in learning: variation in the South of England in the 1980s. *Assessment in Education* 1(2): 201-222

Madaus GF, West MM, Harmon MC, Lomax RG, Viator KA (1992) *The influence of testing on teaching math and science in grades 4-12: Executive Summary*. Boston, US: Centre for the Study of Testing, Evaluation and Educational Policy

Maines B, Robinson G (1996) *B/G Steem: a self- esteem scale with locus of control items*. Bristol: Lucky Duck Publishing

McCombs BL (1999) *Learner-Centred Classroom Practices.* Available from the author. Denver, Colorado, US: University of Denver Research Institute

Neill M, Gayler K (1999) 'Do high stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation tests'. Paper presented at the High Stakes K-12 Testing Conference sponsored by the Civil Rights Project, Harvard University, Teachers College, Columbia University, and Columbia Law School, New York, US (December 1998)

Norwicki S, Strickland B (1973) A locus of control scale for children. *Journal of Consulting and Clinical Psychology* 40: 148-155

OECD (2001) *Knowledge and skills for life: first results from PISA 2000*. Paris: OECD

Osborne M, McNess E, Broadfoot P, Pollard A, Triggs P (2000) *What teachers do: changing policy and practice in primary education*. London: Continuum

Rigby CS, Deci EL, Patrick BC, Ryan RM (1992) Beyond the intrinsic-extrinsic dichotomy : Self-determination in motivation and learning. *Motivation and Emotion* 16: 165-185

Stiggins RJ (2001) *Student-Involved Classroom Assessment.* 3rd Edition. Columbus, Ohio, US: Merrill Prentice Hall

Yeh SS (2001) Tests worth teaching to: constructing state-mandated tests that emphasize critical thinking. *Educational Researcher* 30(9) 2-11

Appendix A: Search strategy

Research studies were identified from the following sources or in the following ways:
Bibliographic databases (ERIC, BEI, PsycLIT, Social Science Citation Index)
Specialist registers (research registers of NFER, SCRE)
Search of journal publishers web pages, or hand searching, of key journals
Personal contacts
Scanning the reference lists of already identified reports
Direct requests to educational research institutions and association members (AEA Europe, AAIA, NFER, QCA)

Search terms for searching bibliographic databases included the following sets in combination:

(i) Terms to indicate that a study is about summative assessment and testing programmes
(ii) Terms to indicate that a study involves pupils aged from 4 to 18
(iii) Terms to indicate that a study concerns motivation

(i)	(ii)	(iii)
Summative	School	Motivation
National test	Primary school	Intrinsic motivation
GCSE	Elementary school	Extrinsic motivation
Certification	Secondary school	Learning motivation
Eleven plus	High school	Self-motivation
Test	Middle school	Attitude
Testing	First school	Learning style
Grading	Pre-school	Deep learning
Examinations	Kindergarten	Surface learning
Public examinations		Learning strategy
Assessment		Learning outcome
Evaluation		Self-regulated learning
Appraisal		Self-efficacy
Key stage tests		Self-esteem
National Curriculum tests		Reward
National Curriculum assessment		Pupil learning
Baseline		Student learning
Portfolio		Learning disposition
Profile		Self-concept
Mandated tests		Mastery orientation
High stakes		Achievement orientation
Accountability		Performance orientation
Transition		Test-anxiety
Transfer		

Searches of electronic databases are complex procedures, which are idiosyncratic to the particular database concerned. Table 1 shows the search strategy used for the ERIC and BEI databases:

Table A1: Search history for ERIC and BEI

#	Search History	Results	Display
1	Evaluation/	2064	Display
2	exp achievement tests/ or exp cognitive tests/ or exp criterion referenced tests/ or exp norm referenced tests/ or exp objective tests/ or exp performance tests/ or exp preschool tests/ or exp standardized tests/ or exp teacher made tests/	9418	Display
3	exp cognitive measurement/	645	Display
4	exp british infant schools/ or exp elementary schools/ or exp high schools/ or exp junior high schools/ or exp middle schools/ or exp nursery schools/ or exp secondary schools/ or exp vocational high schools/ or exp vocational schools/	33407	Display
5	exp learning motivation/ or exp self motivation/ or exp student motivation/	6828	Display
6	Self efficacy/	1416	Display
7	1 or 2 or 3	11891	Display
8	5 or 6	8114	Display
9	4 and 7 and 8	33	Display

Contract

Run Saved Search Save Search History Delete All Searches

Enter **Keyword** or phrase: Map Term to Subject Heading

Perform Search

Searches were made of the contents lists of the following journals, which are the key and leading publications for studies in the area of assessment. The record identifies the dates back to which the journal was searched, the location of the journal (which university library) and the number of articles found.

Table A2: Journals searched with dates and number of articles found

Title of journal	Dates searched	No. of articles
Applied Cognitive Psychology	96-01	0
American Journal of Evaluation	98	1
American Journal of Education	85-93	0
American Psychologist	90-01	0
American Journal of Educational Research	85-01	3
American Research Journal	85-01	0
Assessment in Education	All	5
British Educational Research Journal	All	11
British Journal of Educational Studies	2000-01	1
British Journal of Educational Research	All	
British Journal of Educational Psychology	99-01	4
British Journal of Developmental Psychology	85-01	0
British Journal of Educational Technology	00-01	0
British Journal of Sociology of education	99-01	0
Cambridge Journal of Education	98-01	0
Cognition and Instruction	00-01	0
Cognitive Science	96-01	0
Cognition and Instruction	00-01	0
The Curriculum Journal	00-01	0
Contemporary Educational Psychology	93-01	1
Developmental Review	93-01	0
Educational Studies	99-01	0
Educational Psychology	81-01	2
Educational Psychology Review	97-01	0
Educational Researcher	85-01	2
Educational Research	81-01	1
Educational Review	99-01	0
Educational Assessment	93-01	2
Elementary School Journal	85-01	2
European Journal of Education	99	0
European Journal of Psychology of Education	All	4
Harvard Education review	85-01	0
Evaluation	99-01	0
Forum	All	1
International Journal of Educational Research	95-01	1
Instructional Science	97-01	0
Journal of Curriculum Studies	97-01	0
Journal of Educational Psychology	94-01	2
Journal of Negro Education	64-01	2
Learning and Motivation	93-01	0
Learning and Instruction	95-01	
Oxford Review of Education	99-01	1
Phi Delta Kappan	88-94	0
Psychological Bulletin	All	0
Research Papers in Education	All	1
Review of Educational Research	85-01	4
Research in Education	99-01	0
Sociology of Education	63-95	0
Studies in Educational Evaluation	85-01	1
Teachers College Record	95-01	3
Westminster Studies in Education	78-94	0

The codes used for inclusion and exclusion criteria are listed in Table A3.

Table A3: Codes used in labelling included and excluded studies

Stage of process	**Code**	**Meaning**
DB1 Source	HSJ	Handsearch journals
	JOL	Journals on line contents/abstracts
	Pers	personal
	Ref	Ref from other article
	ERIC	ERIC
	BEI	BEI
	SSCI	Social Science Citation Index
	Reg	Regard Database
	ERA	Education Research Abstracts
	COP	Copac
	NFER	NFER
DB1 Include	Get	Obtain full text
	Got	Full text available (hard & soft copy)
	In	Send to DB2
DB1 Exclude	Out A	Language
	Out B	Not summative testing or assessment
	Out C	Not process or outcome evaluations
	Out D	Not 4-18 School pre-school
	Out E	Not Motivation
DB BR	BR	Background & recommendations
DB2	Got	Full text available (hard & soft copy)
	Not Got	Full Text Unavailable
Include	Key-words	
Exclude	Out A	Language
	Out B	Not summative testing or assessment
	Out C	Not process or outcome evaluations
	Out D	Not 4-18 School pre-school
	Out E	Not motivation
BD3	Map	New criteria from Review Group
DB4	Eppi Reviewer	

Appendix B: Keywords

1. Type of printed material

Primary report
Secondary report
Resource
Policy document

2. Identification of report

Citation
Contact
Handsearch
Unknown
Electronic database
(*please specify*)
……

3. Status

Published
In press
Unpublished

4. Language ***(please specify)***

……

5. Programme Name

…………
……..…

6. Which type of study does this report describe?

A. Outcome evaluation
(i) RCT
(ii) Trial
(iii) Pre and post test
(iv) Post test
(v) Reversal design
(vi) Cohort study
(vii) Case control study
(viii) Other design
B. Process evaluation
C. Economic evaluation
D. Intervention description
E. Methods
(i) Instrument design
(ii) Other
F. Needs assessment
G. Review
(i) Systematic
(ii) Non-systematic
(iii) Meta-analysis
H. Descriptive study

7. In which country/ countries was the study carried out?

……
….

8. What is the topic focus of the study?

Curriculum*****
Disciplines
Methodology
Policy
Organization
Teacher careers
Teaching and learning

Other

*****Curriculum

Art	***History***
Business Studies	***Languages***
Citizenship	***Maths***
Cross-curricular	***Music***
Design & Technology	***PSE***
English	***Phys. Ed.***
Environment	***Religious***
General	***Science***
Geography	***Vocational***
Hidden	***Other ……***

9. What is the educational setting of the study?

Adult education
Community based
Correctional institution
Further education
Government department
Higher education
Home
Initial teacher training
Informal education
Nursery education
Primary education
Pupil referral unit
Secondary education
Other educational body
Workplace

10a. What is the population focus of the study?

Preschool children
Primary children
Secondary children
Post compulsory learners (17- 20)
Adult learners (21+)
Head
Senior management
Teaching staff
Non-teaching staff
Parents
Governors
Local education authority officers
Government
Other education practitioners

10b. Sex of population

Female
Male
Mixed sex

11. Intervention provider
(only for study types keyworded as 6A, 6B, 6C and/or 6D)

Advisor
Community worker
Computer
Counsellor
Examination board
Government
Head teacher/Principal
Health professional
Health promotion practitioner
Induction pack
Inspector
Lawyer
Lay therapist
Local education authority
Parent
Peer
Psychologist
Researcher
Residential worker
Social worker
Teacher/lecturer

12. Type (s) of intervention
(only for study types keyworded as 6A, 6B, 6C and/or 6D)

Advice
Anger management
Counselling
Curriculum
Daycare
Environmental modification
Examinations
Family Therapy
Feedback
Funding
Incentives
Inspection
Instruction
Legislation/regluation
Parent training
Professional training
Rehabilitation
Resource access
Sanctions
Screening
Service access
Skill development
Social support
Staff ratios
Treatment

PTO

13. Purpose of assessment

i. summative
ii. formative
iii. monitoring

14. Assessment agent

i. teacher
ii. internal
iii. external
iv. self

15. Assessment form

v. written only
vi. performance
vii. course work
viii. other

16. Types of learning motivation

viii. Intrinsic Motivation or Extrinsic Motivation or Achievement Motivation or Performance learning or Mastery learning
ix. Self-esteem/Self-efficacy
x. Locus of control/Executive control/ Self-regulated learning
xi. Learning to learn/learning profile/learning journey
xii. Test Anxiety/stress/phobia
xiii. Learning for meaning
xiv. Learning dispositions

Appendix C: Summary of extracted studies

Benmansour N (1999) Motivational orientations, self-efficacy, anxiety and strategy use in learning high school mathematics in Morocco. *Mediterranean Journal of Educational Studies* 4(1): 1-15

Brookhart SM, DeVoge J (1999) Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education* 12(4): 409-425

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

Davies J, Brember I (1998) National Curriculum testing and self-esteem in year 2 – the first five years: a cross-sectional study. *Educational Psychology* 18(4): 365-375

Davies J, Brember I (1999) Reading and mathematics attainments and self-esteem in years 2 and 6 – an eight year cross-sectional study. *Educational Studies.* 25(2): 145-157

Duckworth K, Fielding G, Shaughnessey J (1986) *The relationship of high school teachers' class testing practices to students' feelings of efficacy and efforts to study*. Oregon, US

Evans E, Engelberg R (1988) Students' perceptions of school grading. *Journal of Research and Development in Education* 21(2): 44-54

Ferguson C, Francos J (1979) Motivation and mode: an attempt to measure the attitudes of 'O' level GCE candidates to English language. *Educational Studies.* 5(3): 231-239

Gordon S, Reese M (1997) High-stakes testing: worth the price? *Journal of School Leadership* 7: 345-368

Hughes B, Sullivan H, Beaird J (1986) Continuing motivation of boys and girls under differing evaluation conditions and achievement levels.*American Educational Research Journal* 23(4): 660-667

Johnston J, McClune W (2000) Selection Project SEL 5.1: pupil motivation and attitudes: self-esteem, locus of control, learning disposition and the impact of selection on teaching and learning. Belfast: Queen's University

Leonard M, Davey C (2001) *Thoughts on the 11 Plus*. Belfast: Save the Children Fund

Little A (1994) Types of assessment and interest in learning: variation in the South of England in the 1980s. *Assessment in Education* 1(2): 201-222

Paris SG, Lawton TA, Turner JC, Roth JL (1991) A developmental perspective on standardised achievement testing. *Educational Researcher* 20(5): 12-20

Perry NE (1998) Young children's self-regulated learning and contexts that support it. *Journal of Educational Psychology* 90(4): 715-729

Pollard A, Triggs P, Broadfoot P, Mcness E, Osborn M (2000) *What pupils say: changing policy and practice in primary education.* Chapters 7 and 10. London: Continuum

Reay D, Wiliam D (1999) 'I'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal* 25(3): 343-354

Roderick M, Engel M (2001) The grasshopper and the ant: motivational responses of low achieving pupils to high stakes testing. *Educational Evaluation and Policy Analysis* 23(3): 197-228

Schunk D (1996) Goal and self-evaluative influencs during children's cognitive skill learning. *American Educational Resarch Journal* 33(2): 359-382

How I perceive my capacity to undertake the task				
What I feel and think about myself in relation to learning				
The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Benmansour N (1999) Motivational orientations, self-efficacy, anxiety and strategy use in learning high school mathematics in Morocco *Mediterranean Journal of Educational Studies,* Vol. 4(1) 1-15	Orientation of effort (intrinsic or extrinsic) Self-efficacy Test anxiety Learning strategy	Regular school experiences in maths	Outcome evaluation Post test	High

Aims

To explore students' perceived motivational orientations, self-efficacy, test anxiety and strategy use in mathematics.

Research design

Use of a 36-item self-report questionnaire designed to measure motivational goal orientation, self-efficacy and test anxiety in exploring relations between these characteristics and their variation with sex within regular school experiences within mathematics.

Data collection and analysis

Data were collected through a self-report questionnaire. Factor analysis using Principal Components analysis and a varimax rotation was applied to the items testing motivational orientation and, separately, to the six self-efficacy items, and to the six test anxiety items. The level of endorsement of each of the four emerging kinds of orientation was calculated from the average score for items loading on each factor. Mean scores for the various variables were computed and a series of T-tests applied to test for differences between sexes. Pearson's correlations were computed between motivational orientation, test anxiety and strategy scores.

Author's findings

Four factors were found in the measure of goal orientation:

- intrinsic (desire to learn mathematics out of interest)
- grades (getting good grades)
- social status (making effort to ensure university entrance and high post in the future)
- pleasing others (effort made in order to please the teacher, or parents, etc.)

Students showed the highest endorsement for 'grades' and social status. Less for 'intrinsic' and weakest of all for 'pleasing others'. The items for self-efficacy formed one scale, as did those for test anxiety.

Test anxiety and self-efficacy were negatively correlated with one another, even though the mean scores were about the same.

Benmansour N (1999) cont'd

In terms of frequency of use of active and passive strategies, passive strategies were by far more frequently used. Intrinsic goal orientation was related to both types of strategies and showed a particularly strong association with active learning strategies. 'This suggests that intrinsically motivated students tended to use both types of strategy but were more likely than extrinsically oriented students to make use of active strategies.'

A stronger orientation towards grades was related to high levels of test anxiety and greater use of passive strategies. This implies that working under pressure to achieve high grades may be anxiety provoking and detrimental to students study skills and performances. Intrinsic motivation appears to be the most desirable goal.

Goal orientations and test anxiety: Stronger intrinsic orientation showed a negative relation with test anxiety, whereas the extrinsic dimension of 'grades' exhibited a positive relationship. No relationship between scores on social status or pleasing others and test anxiety. Thus intrinsically motivated students were less likely to experience test anxiety than the extrinsically motivated students were. Students driven by grades were more likely to exhibit test anxiety.

Girls scored more highly than boys did on perceived test anxiety. In comparison with girls, boys tended to show more interest in math, perceived themselves as more capable, less test anxious, and reported using a wider repertoire of strategies.

Author's conclusions

Self-efficacy was related to higher intrinsic orientations, lower test anxiety and use of a wider repertoire of strategies including active ones. This confirms the findings of Pintrich and DeGroot. However, in contrast, where Pintrich and DeGroot found no significant relation between test anxiety and strategy use, this study revealed that test anxiety was negatively related to active strategies – anxious student were less likely to use active strategies. Thus, 'for this sample of students, elevation of test anxiety during evaluative situations may be accounted for by a deficit in their study skills'. 'The results of this paper all point to the conclusion that an emphasis on evaluation may promote extrinsic goals in students, induce higher levels of test anxiety in them, decreases their strength of self-efficacy and inhibit their use of effective strategies.... In order to counterbalance the emphasis placed on grades, teachers need to cultivate in students more intrinsic interest and self-efficacy, which are potentially conducive to the use of effective strategies and better performance.'

Reviewers' differences

Some statements border on claiming causal relationships, the direction of which is often ambiguous: for example, 'elevations in test anxiety may be accounted for by a deficit in study skills'. 'Students' strong extrinsic orientations towards grades and 'social status' may have been shaped by an education system which puts great emphasis on evaluation and selectivity'.

How I perceive my capacity to undertake the task				
The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Brookhart S, DeVoge J (1999) Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education* 12(4): 409-425	Self-efficacy Effort	Regular classroom assessments	Outcome evaluation Other design	High

Aims

To test the usefulness of a theoretical framework for interpreting results of assessment events, which include the following variables: level of perceived task characteristics, perceived self efficacy, amount of invested mental effort, achievement and the relations between these.

Research design

Classroom assessment environment is an important part of classroom atmosphere, and is highly related to classroom achievement. The assessment event is complex and requires an understanding of the variables that are hypothesised to have an effect on outcomes and achievement. This study is part of ongoing work to investigate the classroom assessment environment as having an impact on what students perceive as important to learn and how good they are at learning. A model has been developed and is tested in this study (see table below). The researchers collected data related to naturalistic assessment events taking place during language arts classes in two third grade classes. The theoretical framework that informed the research design is included at the end of this report. Three dependent variables (perceived self efficacy, perceived task characteristics and amount of invested effort) were evaluated in relation to each other and to the independent variable, the intervention, which was the assessment event naturally occurring in the course of teaching and learning. Four classroom assessment events in each of two classes were studied. To describe the classroom assessment environment from the point of view of an observer, language arts blocks of instructional time were observed on two different occasions. Language arts blocks consisted of reading, spelling, and language arts instruction; each of these used different texts, teachers taught them with different lesson plans, and students were made aware of the transitions between them (e.g. 'Get out your spelling books; it's time to do spelling').

Data collection and analysis

Four different classroom assessment events were selected in each class, in consultation with the teachers. For each event, a pre-survey was administered to the whole class to collect perceptions of perceived task characteristics (PTC) and perceived self-efficacy (PSE) to do the task. A post survey was administered after the assessment but before students received feedback, to collect perceptions of amount of invested mental effort (AIME). Achievement was noted as the score the teacher assigned for student performance on the assessment (i.e. percentage correct). Data were obtained from absentees as they made up their work. Student mobility in and out of the district resulted in a small amount of missing data.
Before each assessment event, four students were selected, in consultation with the teacher, to be interviewed about their perceptions of the assessment. The students to be interviewed for each assessment event were selected to vary by gender (boy or girl) and achievement level (low, middle, high). Different students were interviewed each time to maximise information and to allow the children to share the privilege of talking with the researchers, whom they perceived as classroom visitors.

Brookhart S, DeVoge J (1999) cont'd

Descriptive statistics were the main form of analysis. Means, standard deviations and correlational analyses were used on the quantitative data. Coding of interviews into the categories in the theoretical framework was used on the qualitative data.

Authors' findings

1. The model of the role of classroom assessment in student motivation and achievement (Brookhart, 1997a, 1997b; Figure I) held in general; that is, there were relations among the assessment task as students perceived it, their perceptions of their ability to do the task, their effort, and their achievement. Both quantitative and qualitative data supported this conclusion.
2. Students' self-efficacy judgements about their abilities to do particular classroom assessments were based on previous experiences with similar kinds of classroom assessments. Results of previous spelling tests, for example, were offered as evidence about how students expected to do on the current spelling test. This finding is consistent with the model tested and also with self-efficacy research (Lepper, 1988; Schunk, 1994).
3. The relation between perceived self-efficacy and effort is not a simple one because students who perceived themselves to be so capable that the work would not be a challenge would not expend much effort. Specific prior experience with similar assessments may be necessary before students report investing less effort; that is, evidence that makes them sure they will do well. Students who perceive themselves as more efficacious will also tend to be students who report investing more mental effort in performance on an assessment. An exception might be for students who are performance oriented (Ames & Archer, 1988) and who thus might consider putting forth effort as an end in itself, by which they would be judged. Lack of variability among measures and small class sizes did not permit definitive conclusions about this relationship.

Authors' conclusions

The results suggest that information feedback is crucial to further learning. Judgmental feedback may also influence future learning through students' use of it as evidence of their capability to succeed at a particular kind of assessment. Goal orientations (performance or mastery) are also linked to effort. This line of enquiry should be extended to other classroom assessment environments (not always positive like this one) and other grades.
Teacher's explicit instruction and how they present and treat classroom assessment events affects the way students approach them. When a teacher exhorts a student to work towards a good grade that teacher is on the one hand motivating students and other the other setting up a performance orientation which may decrease motivation. Greater attention should be paid to helping students interpret feedback.

Reviewers' comments

None

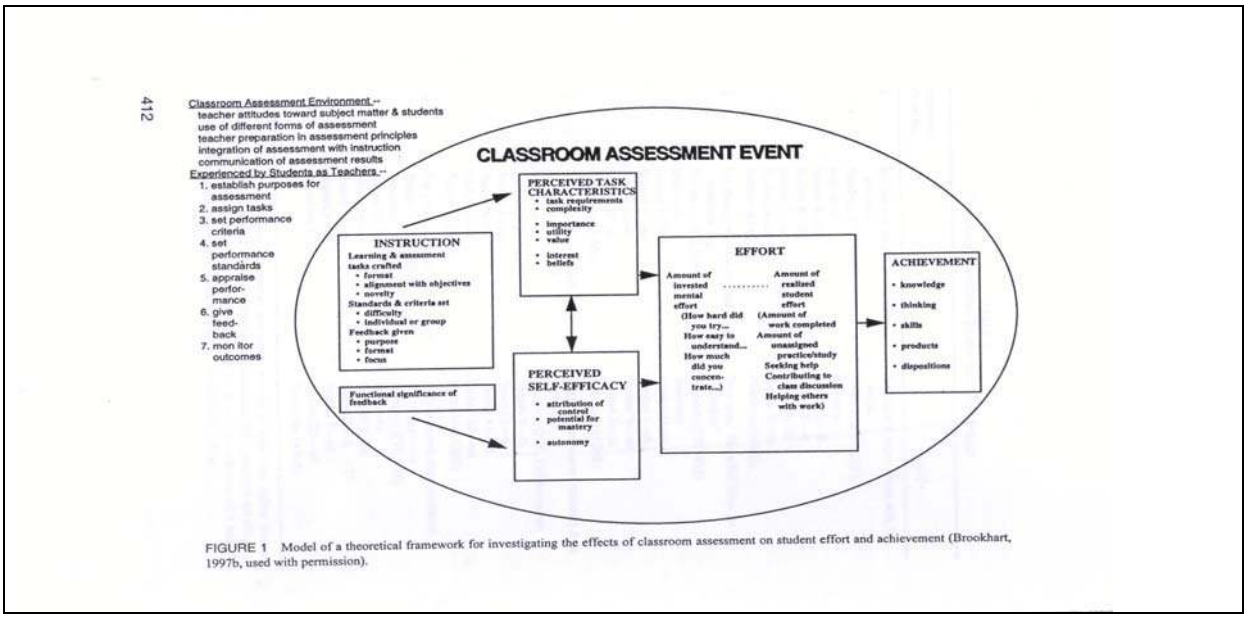


FIGURE 1 Model of a theoretical framework for investigating the effects of classroom assessment on student effort and achievement (Brookhart, 1997b, used with permission).

The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
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	Interest	Type of feedback	Outcome evaluation RCT	High

Aims

To test the hypothesis that intrinsic motivation will be differentially affected by task-involving and ego-involving evaluation and that provision of both kinds of evaluation will promote ego-involvement rather than task-involvement. To test these using convergent and divergent tasks for 5th and 6th graders.

All students were given three work booklets containing the experimental tasks for the three sessions. In each booklet, there were two tasks: one of constructing words from given letters and the other a divergent thinking task. Tasks in sessions 1 and 3 were similar to each other. Tasks in session 2 were slightly different. Students worked on the first task for 10 minutes, then the second for 10 minutes and then answered the interest questionnaire. After sessions 1 and 3 subjects were also asked to state how many additional tasks they would like to receive and after session 3 also to recall the evaluation received on Session 2.

Experimental hypotheses:
(a) Post-test interest and performance on both tasks will be highest after receipt of comment at both levels of school achievement.
(b) High achievers will score similarly on session 2, interest and convergent thinking, in all groups, while low achievers will score highest on immediate divergent thinking after comments.
(c) Subjects who received comments alone will recall these better than subjects who also receive a grade; changes in performance from pre-test to post-test will be related to the content of the comments received earlier in the comments but not in the grades + comments condition.
(d) Patterns of interest and performance on sessions 2 and 3 will be similar in the grades and grades + comments condition.

Research design

The three experimental conditions of feedback were as follows:
1. Comments only: feedback consisted of one sentence, which related specifically to the performance of the individual child
2. Grades: Scores were computed for the word generation tasks by noting the number of short and long words and the total number of words. One point was awarded for short words and 2 points for long words. For the divergent tasks, the number of responses was counted to give a measure of fluency, and counts were made for flexibility, elaborated responses and original responses. Originality was given a score of 2 and other components a score of 1. For the grades group the scores were converted to follow a normal distribution with scores ranging from 40 to 99.
3. Grades + comments group were given both a grade and a comment combining the information given to each of the above groups.

Data collection and analysis

Butler R (1988) cont'd

The sample comprised 132 5th and 6th grade Israeli pupils from 12 randomly selected classes. The instruments consisted of three work booklets containing the experimental tasks for sessions 1, 2 and 3. After each session an interest questionnaire was given. After sessions 1 and 3 subjects were asked to state how many additional tasks they would like to receive and after session three to recall the evaluation they received after session 2. Analysis methods included the scoring of tasks, with mean scores being computed. Two-way ANOVA was used to examine main effects and specific contrasts.

Author's findings

For the convergent tasks, high achievers scored higher in comments conditions and in grades only conditions than in grades plus comments. For low achievers, those in comments conditions scored more highly than those in grades conditions and those in grades only score more highly than grades plus comments. Thus both high and low achievers did better with grades only than grades + comments.
For divergent tasks, those under comments conditions scored more highly than under grades and grades + comments conditions and there was no significant difference between the latter two groups. This was the same for high and low achievers.

Interest is the only outcome measure relevant to this review. High achievers expressed similar interest in all feedback conditions, whilst low achievers expressed most interest after comments. The combined interest of high achievers receiving grades and grades + comments was higher than that of low achievers in these conditions. The interest of high and low achievers in the comments only group did not differ significantly.

Author's conclusions

The results of the study suggest that some of the difficulties faced by means-ends analyses in conceptualising and predicting how interest and performance can be maintained or enhanced, and not just undermined, can be resolved by distinguishing not only between constrained and non-constrained, but also between task-involved and ego-involved task engagement. Different motivational orientations have implications not only for subsequent interest but also for immediate interest and performance and subsequent performance. Combining task and ego-involving evaluation will induce an ego-involving orientation, just as does the provision of ego-involving evaluation alone. This study implies that promoting task involvement may also promote the interest and performance of most students.

Reviewers' comments

None

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Davies J, Brember I (1998) National Curriculum Testing and self-esteem in year 2 – the first five years: a cross-sectional study. *Educational Psychology* 18(4): 365-375	Self-esteem	National curriculum tests	Outcome evaluation Case control	High

Aims

To report the changes in different, but equivalent, cohorts of year 2 children over the period during which the National Curriculum and National Curriculum Tests were introduced. The changes were measured in self-esteem, maths and reading achievements.

Research design

The administration of measures of self esteem, reading and maths to year 2 children in the years running up to the introduction of the tests at KS1 and for three years following the introduction. Similar measures were applied to cohorts of year 6 children for the same five years, prior to the introduction of KS2 tests. Thus any impact of the tests on self-esteem and achievement could be investigated for year 2 children and compared with the self-esteem of year 6 children.

Data collection and analysis

Class teachers administered both the Primary Reading Test and the NFER Maths Tests to children in the cohort after their half-term holiday in the summer term. Researchers administered the Lawseq questionnaire as a measure of self-esteem. The sample comprised five cohorts of year 2 children, (213, 209, 216, 204, 216) totalling 1058 and five cohorts of year 6 children (176, 207, 199, 220, 196) totalling 998.
Means of self-esteem measures were calculated for all cohorts. A two-way analysis of variance was used (with harmonic means to compensate for the uneven cell numbers) with cohort and year group as independent variable.
Analyses of variance were used for each age group to examine difference between age groups. Difference between means for each year group were computed and examined for statistical significance.
Maths and reading performance means were also calculated for each cohort and year.
Correlations between self-esteem and attainment scores in maths and reading were also computed.
For one cohort only (the first, which had been tested both at year 2 before the National CurriculumTests), self-esteem in year 6 could be compared with self-esteem in year 2.

Authors' findings

The administration of measures of self-esteem, reading and maths to year 2 children in the years running up to the introduction of the tests at KS1 and for three years following the introduction. Similar measures were applied to cohorts of year 6 children for the same five years, prior to the introduction of KS2 tests. Thus any impact of the tests on self esteem and achievement could be investigated for year 2 children and compared with the self esteem of year 6 children.

For year 2 children, self-esteem dropped with each year, with the greatest drop coinciding with the introduction of the National Curriculum Tests. Although there was a small upturn for the fifth cohort, the level still remained lower than the third and very much below the second cohort. Means for the pre-national test cohorts were significantly higher than for the post-national test cohorts.

Davies J, Brember I (1998) cont'd

The difference in self-esteem across cohorts was highly significant for year 2 children but not for year 6 children.

There was no overall relationship between drop in self-esteem and achievement in reading and maths on the standardised tests. However, there was a positive correlation between self-esteem and performance after the introduction of National Curriculum Tests (although these were the children with the lower self-esteem).

The cohort who was tested as year 2 and again as year 6 showed very little change in self-esteem after four years.

Authors' conclusions

The drop in self-esteem occurred only in year 2 children after the introduction of the tests. It was not found at all in the year 6 children and so there was no general drop in self-esteem. The lack of correlation between achievement and self-esteem before the National CurriculumTests means that 'the children's view of themselves was apparently less affected by their attainments than the post national test group'.

The authors suggest two aspects of the tests that could have affected the children's self-esteem: their use as a summative evaluation and their administration. In relation to the first, the parents aspiration for the children to do well and the teachers' concern about being held responsible for the children's performance would put pressure on the child. In relation to the second, the teacher was involved in giving attention to individuals (see above) for considerable time during these tests, which could make other children feel undervalued. Moreover they suggest from other evidence that the teachers' morale was low at this time due to the tests and this could have communicated itself to the children.

However, the authors note that tests may be just one factor affecting self-esteem. They suggest that the administration process of the tests may have affected children's self-esteem, as this affects the day-to-day exchanges between teacher and pupils and so could mediate the effects on self-esteem.

Reviewers' comments

The size of the correlations is very small and educationally insignificant. However, this does not affect the main findings.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Davies J, Brember I (1999) Reading and mathematics attainments and self-esteem in years 2 and 6 – an eight year cross-sectional study *Educational Studies* 25 (2): 145-156	Self-esteem	National curriculum tests	Outcome evaluation Case control	High

Aims

To measure self-esteem, mathematical attainment and reading attainments for a series of 8 cohorts of year 2 and year 6 children spanning the period 1989 to 1997, that is, from one year before the national curriculum tests were introduced. 'Self-esteem and the effects of high stakes national testing is an area of enquiry for the National Curriculum.' Also implicitly to study the extent to which standards of achievement were raised after the introduction of the National Curriculum.

Research design

The study administered tests and questionnaires to eight cohorts of year 2 and year 6 children in each of 8 successive years 1988 to 1995. Measures made were:

- self-esteem as measured by the Lawseq questionnaire (Lawrence 1982)
- maths attainment using NFER tests
- reading attainment using the Primary Reading Test (France 1981)

The first two year 2 cohorts and the first six year 6 cohorts were tested before the introduction of tests to their age group. Thus changes in the measured features associated with the introduction of the National Curriculum Tests could be investigated.

Data collection and analysis

The cohorts consisted of 1,513 year 2 children (cohorts of between 178 and 198) over 8 years and 1,488 year 6 children (cohorts of 160 and 207) over 8 years. Means of the Lawseq scores were computed for each cohort for each year group. Two-way analyses of variance were computed for self-esteem scores with cohort and year group as independent variables (harmonic means were used to compensate for the differences between class sizes). One way analysis of variance was carried out with cohort as the independent variable followed by tests of significance between cohort scores. Correlations between the Lawseq scores and performance on maths and reading were computed. National test scores for the last two cohorts of year 6 were correlated with self-esteem.

Authors' findings

There were significantly different patterns in the changes in self-esteem scores for the year 2 and year 6 cohorts. Within each year the difference between cohorts was significant.
For year 2, means of self-esteem dropped significantly for cohorts 3 and 4 (year 3 coinciding with the start of NC testing) but recovered gradually so that for cohort 8 the scores were significantly higher than for cohorts 4 and 5.
For year 6, 'the means for cohort 8 were significantly higher than Cohorts 1 to 5 and that cohorts 6 and 7 were significantly higher than cohort 2'.
The correlations for the total sample between Lawseq and attainment scores showed was positive but when pre National Curriculum Tests cohorts in year 2 were taken separately there was no significant correlation, whilst for the post-National Curriculum Tests cohorts the correlation was significant. For year 6 all correlations, pre and post NC test cohorts were significant. Year 6 post NC test cohorts did better than pre NC test cohorts in all test scores.

Davies J, Brember I (1999) cont'd

Authors' conclusions

'The year 2 children's self-esteem dropped significantly when National Curriculum Tests were first introduced and recovered to almost pre National Curriculum tests in the eighth year of national testing' (actually the sixth; author in error here). Year 6 children's self-esteem means fluctuated for the first six years and increased significantly in the final two years after the introduction of National Curriculum Tests.

The authors suggests that the reason that the year 2 children's self-esteem scores do not remain low may be owing to the fact that the initial shock of the new National Curriculum assessment procedures to both teachers and children had lost some of its initial impact. The process of testing changed considerably over the first three years and the amount of time required for administration declined noticeably. Teacher felt inadequate to meet the demands of the first tests and this affected their self-esteem and, through them the children's.

In the case of year 6, the children were not tested until four years after the first KS1 tests and so there was time for an assessment culture to develop in the school. They were also more mature than when first tested in year 2.

Reviewers' comments

The authors seem reluctant to recognise that for year 6 children the introduction of the National Curriculum tests was associated with a rise in self-esteem and in achievement. No interpretation of this finding is offered other than the development of an assessment culture in the school. It may well be, for example, that the tests operationalised some of the goals of learning for teachers who then directed their teaching more effectively. Other changes occurring during the time of the study, such as in the curriculum, are likely to have had an impact.

How I perceive my capacity to undertake the task				
The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Duckworth K, Fielding G, Shaughnessy J (1988) *The relationship of high school teachers' class testing practices to students' feelings of efficacy and efforts to study*. Centre for Educational Policy and Management, Oregon USA.	Self-efficacy Effort	Teachers' grading practices	Outcome evaluation Post test	High

Aims

To develop and test a model of the linkage between high school students' feelings of self-efficacy and efforts to study and high school teachers' testing practices and high school practices across subjects. To understand the relationship between effort, motivation, efficacy and futility in relation to type of teacher feedback so as to inform assessment practice.

Research design

1. Indices of effort, motivation, efficacy and futility measured by questionnaires were developed as a basic model of predictors of students' efforts to study.

2. The relationship between these indices was examined at an individual level. Hypotheses were:
   (i) that the effort to study is a positive function of academic motivation
   (ii) That effort is a positive function of efficacy from obtaining rewards from that effort

3. The relationship between teachers class testing practices and student efficacy, effort and futility was tested in the form of three hypotheses:
   (i) that students clarity of learning objectives would be positively related to student efficacy and effort and to the students experience of teacher communication about what was to be covered in tests and feedback after tests about what students still need to learn
   (ii) that efficacy would be a positive function and futility a negative function of the degree that tests fit what the students studied
   (iii) that students' feelings of futility are a function of the degree of teacher help after students do poorly on tests

4. School and departmental policies: to examine the effect on class testing practices

Data collection and analysis

Descriptive statistics were computed for the questionnaire data, including means and standard deviations. Descriptive correlation analyses of class testing practices and student efficacy, effort and futility were carried out.

Authors' findings

1. Individual level: Motivation and efficacy were statistically related to effort even when controlling for academic aptitude and the negative relationship of futility to effort was weaker and significant on in English. These effects remained after controlling for aptitude but after controlling for tracking of

Duckworth K, Fielding G, Shaughnessy J (1988) cont'd

students in the same classes the relationship between motivation and effort is reduced but the relationship between efficacy and effort remains.
2. Class level: in general the relationships were stronger than found at the individual level
3. Hypotheses re teaching practices:
(i) scattered subject specific results
(ii) only scattered evidence supporting the hypothesis
(iii) varied results between subjects and number of failing students
4. School and departmental policies: No evidence was found that instructional policy at the school or departmental level would promote desirable class testing practices. Nor was there evidence of much collegiate action at the departmental level. But collegiality between teachers was related to student's feelings of efficacy and levels of effort.

Authors' conclusions

Students' feelings of efficacy and futility are functions of the level of clarity regarding test expectations created by teachers' practices in communicating test expectations. Efficacy and futility are functions of the correspondence of tests to those expectations resulting from teachers' practices in constructing tests. Students feelings of futility are a function of the degree of teacher helpfulness after students do poorly on tests. Student feelings of efficacy are a promising mediating variable between teachers class testing practices and students efforts to study.

Students' perceptions about communication, feedback, correspondence, and helpfulness are strongly related to students' feelings of the efficacy versus futility of study and the student feelings of their own effort to study. Authors therefore argue that increasing student perceptions of desirable class testing practices may increase feelings of efficacy and level of effort.

It is possible that the informal culture of expectations built up over the year by teacher remarks and reactions operates independently of the specific practices studied. This may be part of a 'halo' effect from desirable class testing practices.

Collegiality (amount of constructive talk about testing) amongst teachers is related to pupils' perceptions of desirable testing practices, and students' feelings of efficacy and effort. School leadership is needed to develop and foster such collegial interaction.

Teacher collegiality is important and should be encouraged.

Reviewers' comments

The reviewers support the findings and conclusions.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Evans E, Engelberg R (1988) Students' perceptions of school grading. *Journal of Research and Development in Education* 21(2): 45-54	Attitude to grades Attribution	Teachers' grading practices	Outcome evaluation Post test	High

Aims

To test hypotheses about children's reactions to and understanding of grades, derived from previous research and theory in the field of social cognition, cognitive development theory and classroom interaction.

The hypotheses were:

- First, in relation to attitudes to grades, older students are more likely to be critical of grading practices and less accepting of grades received than younger students; older students are more likely to rate grades as important and higher achieving students more likely that lower achieving students to like being graded and see grades as important
- Second, in relation to understanding, older students and higher achieving students are expected to have a better grasp of grading schemes than will younger students
- Third, in relation to attribution, older and higher achieving students are expected to use ability attributions to explain successful grades and higher achievers more likely to attribute successful grades to internal factors and lower achievers to external factors.
- Finally, females more than males are expected to endorse external attributions as causally related to getting good grades.

Research design

This is an exploratory descriptive study, using intact, non-randomised groups. It is a study of how understanding of grades, attitudes to grades and attribution vary with age, achievement and gender. It was carried out through a questionnaire study, using the same instrument for students of grades 4, 6, 7, 8, 9 and 11.

Data collection and analysis

An 88-item questionnaire was administered under standard conditions across all classrooms. Grade level, achievement level and gender composed the independent measures. Achievement level as determined by cumulative grade point average was established using a median level split. Scale scores for attitude, cognitive understanding and attribution constituted the independent measures. Means, ranges and correlations for the three scales were calculated and reported.

Evans E, Engelberg R (1988) cont'd

Authors' findings

On attitudes to grades,

- Significant differences were found in relation to age on four scales. Younger students more often than older ones reported that teachers graded fairly, both for reward and punishment, and also attached less importance to the grades they received.
- Higher achievers (particularly the older students) more than lower achievers saw grades as fair, and liked being graded.
- There was no gender difference in relation to attitudes.

On cognitive understanding of grades:

- Older students knew more about grades, both for simple and more complex schemes, and younger students consistently reported not knowing the answer to questions about grading systems.
- Higher achievers had a better grasp of simple grading schemes and low achievers more than higher reported that grades were influenced by external characteristics of the school situation. No difference by achievement in those who claimed they did not know about complex grading schemes.

No gender difference in relation to cognitive understanding.

Authors' conclusions

The hypothesis that attitudes of older students would differ from those of younger were upheld. But no age difference was found for students' perceptions of the importance attached to grades by their parents.

The hypothesis that older students would have a better understanding of grades was upheld in terms of simple grading schemes. The use by elementary grade teachers of mixed criteria (effort, ipsative, as well as criteria) may confuse the meaning of grades for the younger children. Even older children did not understand complex grades.

The hypothesis relating to attribution was partially upheld. Differences in relation to lower achieving students making more external attributions was upheld, but gender difference were in the opposite direction to that expected (more females making internal attributions). Age differences were as hypothesised even though this conflicted with some other research. This was explained in terms of the greater age range in the present study.

Reviewers' comments

The reviewers agree with the findings and conclusions of the study from the evidence provided.

The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Ferguson C, Francis J (1979) Motivation and mode: an attempt to measure the attitudes of 'O' level GCE candidates to English language. *Educational Studies* 5(3): 231-239	Attitude to subject	Procedures for GCE (Modes 1 and 3)	Outcome evaluation Case control	Medium

Aims

The aim of the study is to explore the disparity in achievement between candidates entered for GCE mode 1 (examination only) and candidates entered for GCE mode 3 (continuous assessment and in-course assessment by teacher). Specifically the study investigates one of the major O level subjects (English Language) in an attempt to determine the inter-relationships between attitude to English Language as a subject, the method of assessment (traditional or mode 3) and attainment in the subject as measured by the grade awarded by the board.

Research design

Five aspects of English formed a theoretical basis for the production of a questionnaire containing 42 statements drawn from a list of 120 statements. Two questionnaires were piloted with 100 'O' level English candidates. Factor analysis was computed on the data from the questionnaire which led to the identification of three concepts:

- enjoyment of English lessons
- enjoyment and value of English through reading and literature
- value of English as against other subjects

Thirty statements from these questionnaires were used for the final version, which was administered to a sample of 792 'O' Level English candidates drawn from 16 centres representing both modes of examination.

Data collection and analysis

Factor analysis was performed on the data from this final questionnaire. The sample of candidates was divided into eight subgroups according to mode, gender and place of study. A discriminant analysis was carried out on the attitude test scores. The analysis was designed to extract the statements that discriminated most among the eight groups. Correlations between attitude score and examination grade were calculated for the eight groups.

Authors' findings

There are significant differences between gender and place of study (i.e. school or college) in relation to attitude and attainment but not between modes of examination (i.e. course work assessment or examination assessment).

Authors' conclusions

It is not possible to explain the differences in success between mode 1 and mode 3 candidates on the basis of attitude.

Reviewers' comments

There is broad agreement with the findings.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Gordon S, Reese M (1997) High stakes testing: worth the price? *Journal of School Leadership* 7: 345-368	Self-esteem Test anxiety	State mandated tests (Texas)	Descriptive study	Medium

Aims

To gather in depth qualitative data on the perceptions of teachers in Texas public schools regarding the effects of TAAS (The Texas Assessment of Academic Skills). It aims to look at the effects of this testing on 'curriculum, teacher decision-making, instruction, student learning, school climate, and teacher and student self-concept and motivation'. The mandated TAAS tests were introduced for the same reason as other state tests (i.e. as 'the primary vehicle for attempting school reform').

In particular, the study sought to describe participants' perceptions of (a) how students are prepared for TAAS (b) effects of TAAS on students (c) effects of TAAS on teachers and (d) the effects of TAAS on students.
Note: The data extraction here is focused on (b) and (c).

Research design

The study probed the views and experience of teachers from all kinds of schools on the use and effects of the TAAS. Themes were identified in the written responses to four entirely open-ended questions:

1. How are students at your school prepared for TAAS?
2. What are the effects of TAAS on your students?
3. What are the effects of TAAS on you as a teacher?
4. What are the effects of TAAS on your school?

These themes were then explored in in-depth interviews with 20 of the original 100 teachers. These were randomly selected with 10 from high achieving schools and 10 from low achieving schools.

Data collection and analysis

The authors state that the individual in depth interviews were the primary data-collection method. Interviews were tape-recorded. For each of the main four questions sub-topics were covered, as in the analysis of the survey data. For example, for the effects on students, the sub-topics were (a) emotional effects, if any; (b) academic effects, if any; and (c) social effects, if any.
For the interview data, interview transcripts were coded, categories formed, and data displayed on matrices using the same process followed in the analysis of the survey data: responses were noted for each theme against the combination for variables relating to the school of the teacher responding.
Interview themes were defined as 'a perception reported by at least 16 of the 20 interviewees'.
The interview themes are listed with the numbers of interviewees agreeing with the theme. Each of the four main questions is then considered and trends reported with examples quoted from the interview transcripts.

Authors' findings

There were few differences in the responses of the teachers from high and from low achieving schools.
The 'themes' for effects on students were:

Gordon S, Reese M (1997) cont'd

- a wide range of emotional effects
- a lowering of self esteem of at risk students
- improvements in test-taking skills and TAAS scores

For effects of teachers, the themes were:

- more emphasis on TAAS related content; less emphasis on other content
- more direct instruction and dril1 4 to 8 weeks prior to TAAS
- stress increases as TAAS approaches
- concern, frustration, and disappointment when at risk students perform only poorly on TAAS
- accountable for TAAS related content but not accountable for effective teaching

Emotional effects of students were identified as of three kinds:

- no effect (student who don't realise the importance of TAAS)
- moderate effect (which tends to motivate them to work harder)
- high levels of stress (for both high and low achievers) which leads to anxiety and panic, or anger and resentment, or 'shutting down' (when they tell themselves they have no chance of doing well)

The strongest emotional effects concerned at risk students. 'According to 17 interviewees, failure lowers the already poor self-esteem of these students.'

When asked if they could identify any new learning taking place as a result of the presence of TAAS, 15 (or 20) teachers reported that 'the only new things students had learned were test-taking skills in general and the TAAS format in particular'.

Gains in TAAS scores were ascribed to improved test-taking skills.

13 teachers conjectured that, overall, students may actually learn less academic content as a result of TAAS. Important learning is neglected if it is not TAAS.

According to several interviewees, the TAAS effect of reducing the scope and depth of student learning was especially harmful to higher ability students.

Perceived effects on schools:

- A lot of school resources put in for TAAS materials and staff development even when the budget was tight.
- Non-TAAS related parts of the curriculum de-emphasised.
- No effect on school climate except for the last few weeks before TAAS taken.
- 10 teachers reported overt competition between schools. TAAS scores were a source of pride to parents and the community and therefore brought pressure on the school for high scores. But teachers, especially those from lower SES schools, felt that TAAS should not be used to evaluate the quality of a school.
-

Authors' conclusions

The authors conclude that TAAS results in teaching to the test and the neglect of many aspects of the curriculum. 'We fear that in many of the respondents' schools, curriculum goals like the development of self concept, ethical values, social skills and an understanding of the appreciation of diverse cultures are in danger of being discarded.'

Teachers are adopting procedures which are effective in preparing students to pass the tests, but they can to this 'even though the students have never learned the concepts on which they are being tested. As teachers become adept at this process, they can even teach students how to answer correctly test items intended to measure students' ability to apply, or synthesise, even though the students have not developed application, analysis or synthesis skills'.

'We are concerned that the repetitive demonstrations, drills, worksheets and practice tests may lower the motivation, curiosity and cognitive growth of both teachers and students.'

They concluded that high-stakes testing has negative effects on curriculum, teacher decision-making, instructions, students learning, school climate, and teacher and students' self-concept and motivation. TAAS results in teaching to the test and test format, as the expenses of large parts of the curriculum. personal and social development is neglected.

Reviewers' comments

No differences between reviewers and authors, but perhaps the reviewers would not want to generalise in such a definitive matter.

The Energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Hughes B, Sullivan H, Beaird J (1986) Continuing motivations of boys and girls under differing evaluation conditions and achievement levels. *American Educational Research Journal* 23(4): 660-667	Interest Effort	Experimental word identification and creations exercise	Outcome evaluation RCT	Low

Aims

The aims of the study is to investigate the impact of task difficulty level, source of evaluation and sex of subject on continuing motivation to return to task.

Research design

Two sets of each of two versions of a word-search activity and word identification, one easy and one difficult, were created. Subjects were blocked by sex and randomly assigned to one of the four treatment groups defined by crossing the two difficulty levels with the two valuation conditions. Trained PhD candidates administered treatments simultaneously in separate rooms. Teacher evaluation subjects were told, 'Your scores will be given to your teachers. Your teacher will evaluate your performance'. Self-evaluation subjects were told, 'Your score will not be given to your teacher. Your performance will not be evaluated in any way'. After collecting the papers, the experimenters gave subjects the answers to the items and the option of working on another word-search activity or on the task to create words, emphasising that neither would be collected. Both activities were distributed to subjects. Subjects selected one activity. Experimenters collected the other. Measurements of motivation were the uptake of option to undertake further word-search activities.

Data collection and analysis

Two hundred predominantly Anglo 5th grade students from three suburban elementary schools participated in the study. A 2 (evaluation and condition) x 2 (task difficulty level) x 2 (sex) factorial design was used. The dichotomous motivation scores of subjects returning to a word search activity were analysed with ANOVA.

Authors' findings

There was a higher rate of return to task for students receiving teacher evaluation. The authors warn that caution must be exercised in generalising to longer-term motivational patterns because of lack of sustained effect in delayed measure. Boys show higher rate of return to task than girls do.

Authors' conclusions

This research does not provide support for the conclusion of some authors that grading may reduce continuing motivation. It produces support for researchers who have proposed that performance relevant rewards do not necessarily result in detrimental effects on motivation. Return to task is highly subject to situational variables that vary among classrooms and schools.

Reviewers' comments

The reviewers do not agree with the claims to contribute to motivation theory on the basis of this study. No data presented for the actual tasks and questions regarding validity of measures as measures of motivation. Unclear what sort of motivation are actually being measured and how task specific these measures of motivation are. No data presented about actual tasks and questions regarding validity of measures in relation to motivation.

How I perceive my capacity to undertake the task				
What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Johnston J, McLune W (2000) Selection projects SEL 5.1: Pupil motivations and attitudes – Self-esteem, locus of control, learning dispositions and the impact of selection on teaching and learning. In *The effects of the selective system on secondary education I Northern Ireland. Research papers 2*. Bangor: Department of Education	Learning dispositions (preferences) Locus of control Self-esteem	Classroom climate created by 11+ testing in NI	Outcome evaluation Other design	High

Aims

To investigate the impact of the Northern Ireland Selection for secondary education on teachers, pupils and teaching and learning processes in science lessons.
The study is part of an extensive review of the selection system of Northern Ireland initiated by the Minister of Education in order that decisions on the future of the system should be based on informed discussion and debate.

Research design

Pupils and teachers in 63 year 6 classrooms in a structured opportunity sample of 22 primary schools across Northern Ireland were interviewed and completed questionnaires. The naturalistic intervention was the 11+ examination that is taken in the first term of year 7. The evaluation measured the learning dispositions, locus of control and self-esteem of girls and boys in P6 and again in P7 and related these to the transfer grades obtained by the pupils in the 11+ examination.
The study was underpinned by a conceptualisation of learning expressed in the Interactive Learning Model (see below). It represents learning as a complex process in which 'individual learning combinations of knowledge and understanding, manner of performing and sense of self as a learner occur in identifiable patterns or dispositions'. Four main learning dispositions were identified:

- 'precise processing' (preference for gathering, processing and utilising lots of data, which gives rise to asking and answering many questions and a preference for demonstrating learning through writing answers and factual reports)
- 'sequential processing' (preference for clear and explicit directions in approaching learning tasks)
- 'technical processing' (preference for hands on experience and problem solving tasks; willingness to take risks); technical processors tend to be creative, private and non-verbal thinkers
- 'confluent processing (typical of creative and imaginative thinkers, who think in terms of connections and links between ideas and phenomena and like to see the 'bigger picture')

Instruments were selected to measure learning disposition, locus of control and self-esteem. Classes were observed during normal science lessons.

Johnston J, McLune W (2000) cont'd

Data collection and analysis

The instruments used for measuring self-esteem, locus of control and learning orientation were previously published and validated measures. These were the Learning Combination Inventory (Johnston 1996), the B/G steem scale for primary pupils (Maines and Robinson 1996) and the Locus of Control Scale for Children (Norwicki-Strickland 1973).The Learning Combination Inventory (LCI) was adapted for the purpose of the study and factor analysis confirmed the existence of three factors relating to learning disposition: precise/sequential; technical; and confluent. It was administered to the pupils by a researcher whilst another researcher interviewed the teacher. The teachers later administered the self-esteem, locus of control and attitudes to science questionnaires to their own pupils. Focus group discussions were conducted with pupils on a two-group per class basis. The discussions were taped and transcribed, and analysed using a qualitative data computer package. Patterns were explored among the measured pupil variables and the qualitative data were analysed for perceptions of teachers and pupils as to the effect of the transfer tests on teaching and learning.

Authors' findings

Learning dispositions of children showed a considerable preference, across all transfer grades for 'technical processing'. There was a highly significant difference between this and the other two learning dispositions (precise/sequential processing and confluent processing).
There was significant positive correlation between a precise/sequential learning disposition and self-esteem. There was also a positive correlation between precise/sequential processing and 'enjoyment of science' and between self-esteem and locus of control.

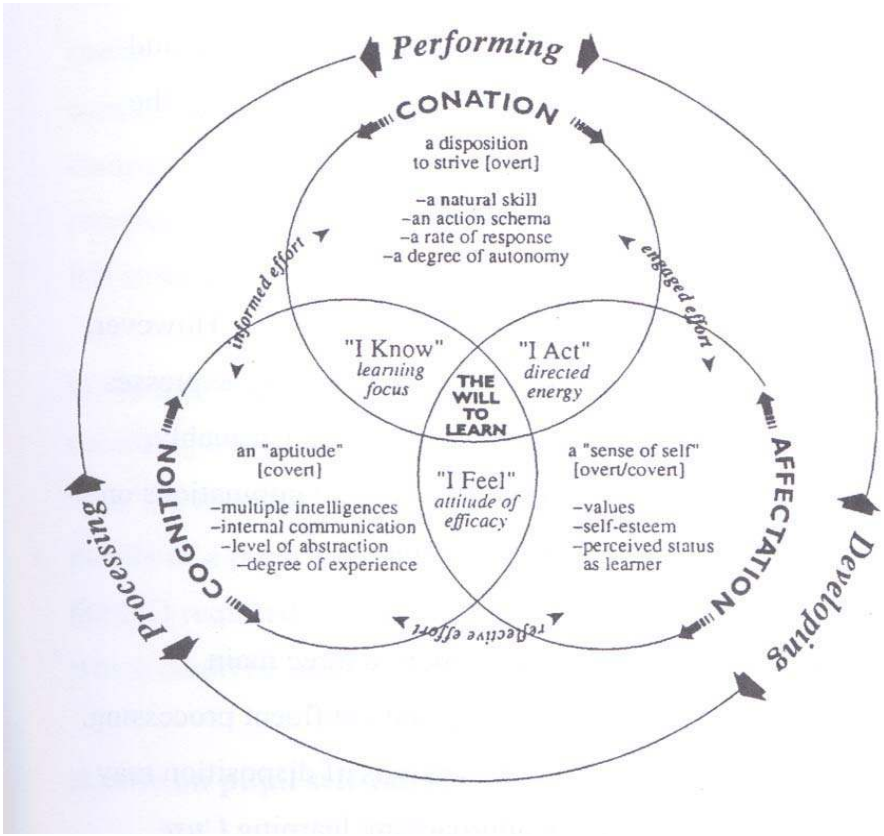
Girls were significantly less favourably disposed towards technical processing than boys. Girls in transfer grades A B1 and B2 were the only ones to report using technical processing as a 'use-first' disposition. Girls were significantly more inclined than boys to use precise/sequential processing and less inclined (particularly those in transfer grades C and below) to use technical processing.

The more positive a pupil's disposition towards precise sequential or technical processing the higher their self-esteem and the more internal their locus of control. Conversely the more confluent a pupils learning orientation the more external their locus of control and the lower their self-esteem.

Boys and girls report significantly different levels of self-esteem.
There are no significant differences between boys and girls across the sample on locus of control, but when analysed by transfer grade group significant differences are found.

Observation of teaching indicated that teachers teach in ways that give priority to sequential processing and link success and ability in science to precise/sequential processing learner characteristics.
Interviews with teachers indicated that they felt the need to teach in this way (through highly structured activities and transmission of information) on account of the nature of the selection tests. 'Where this happens, teachers may (inadvertently or otherwise) value the precise/sequential processing approaches to learning more than other approaches' and in so doing discriminate against and demoralise children who do not learn in this way.

Johnston J, McLune W (2000) cont'd
Authors' conclusions
Teachers perceive the existence of selection and transfer tests to impact significantly on teaching and learning in the primary classroom. This produces teaching and learning which is heavily focused on transmission of factual knowledge. There is therefore a reduction in value placed upon experiential learning and conceptual understanding during the period of the selection process. This means that some children have experiences that fail to meet their needs as learners and leave them demotivated, uninterested and misunderstood as learners. Pupil performance in transfer tests is inextricably linked with pupil learning disposition, pupil self-esteem and pupil locus of control. Boys and girls have significantly different dispositions as learners.
Reviewers' comments
No significant differences



The Interactive Learning Model

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Leonard M, Davey C (2001) *Thoughts on the 11 plus*. Belfast: Save the Children	Attitude to test Self-esteem Impact of tests on teaching and learning	11+ Tests in Northern Ireland	Process evaluation	High

Aims

The aim of the research study was to provide a child centred perspective on the experience of taking the transfer test for selection to secondary education in Northern Ireland.

Research design

The sample for the study was one primary 7 class in each of eight schools. Children were interviewed in focus groups on three occasions between December 2000 and February 2001. The first interviews were conducted just after the children had taken the second transfer test. The second was carried out in the week before the results were announced and the third a week later after the results had been announced. Children in all P7 classes in the schools were also requested to write a story, just after taking the tests, entitled 'My Experience of the 11 plus'. Between the second and third interviews, children in all the classes were invited to draw a picture reflecting their thoughts on the 11 plus.

Data collection and analysis

The study aimed to elicit the views and experiences of children about transfer tests. Focus groups were used because they give children more power to set the terms of discussion. Researchers asked children to describe events from their daily lives as a means of drawing out their subjective experiences. Stories were selected as a second methodology because they accommodate a child's medium of thinking, writing and talking, thus gaining maximum insight into their experiences of the transfer exam. Drawings, as evidence, were selected because they provide additional means for children to express feelings and opinions, which may not be readily communicated in words, or in the other forms of data collection used in the study.

The data comprised the stories, recorded interviews from focus groups and teachers and the children's drawings. No attempt was made to analyse the interview data quantitatively. The bulk of the report is an account of those views and opinions, organised around the research questions:

- preparing for the 11 plus exam
- deciding to do the 11 plus exam
- children's experiences of the 11 plus
- waiting for the 11 plus results
- publication of 11 plus results
- children's thoughts on the 11 plus

Under these headings, the findings are set out in general terms illustrated by quotations.

The 193 drawings were analysed for the themes depicted. 12 themes were picked out as the main ones.

Themes from the children's stories (written after taking the tests but before results were published) were combined with the material from the focus groups in descriptive accounts for the main themes of the report. Children's responses in the two phases were grouped according to the expected grade they would achieve. In the third phase, the actual grade obtained was used in reporting statements.

Leonard M, Davey C (2001) cont'd

Authors' findings

At first, some children found the weekly preparation for the test novel and exciting, but this gradually turned to boredom. Children who experienced high scores in the practice tests experienced a surge in confidence, while those who experienced low scores experienced low self-worth. By the time for decision-making about whether to take the test, some children knew they were likely to fail but still wanted to take the test; others opted out.

Given the practice, most children assumed they would take the test. Some did not want to but felt obligated to respond positively to parental wishes. Some parents used incentives, others enlisted tutors. Lower ability children appeared to play a more active role in decision-making. Children perceived teachers as playing a minor role in decision-making.

The majority of the children approached the tests with fear and uncertainty. On the day of the test, the majority were overwhelmed by the knowledge that the tests were 'the real thing'. Apprehension was heightened by the presence of the invigilator and the formal seating arrangements. The children experienced relief once the exams were over, but once they settled down they realised that it was a long wait for the results.
Children perceive the period of waiting for the exam results as long and stressful. A number expressed their constant worries about how well they had performed. Children generally perceived their parents to be supportive, but some felt pressurised by expectations. Having successful older siblings was a significant worry factor for some.

The children's drawings underlined the mainly negative images children hold towards the 11 plus. Only 4 of the 193 pictures could be interpreted as supportive of the tests. 26 children drew images of children crying. 23 showed children doing the practice tests and illustrated their frustration with their heavy workload. 45 pictures showed children taking the tests and highlighted their stress and concern. A number of children had unrealistic expectations about their own performance despite practice (expecting a C or above). The D grade was viewed with disdain by many. A-grade children were perceived as smart while D-grade children were perceived as stupid. These labels were undisputed, despite parental and school efforts to promote different views.

Children who received high grades felt euphoric; children who received low grades felt devastated. Many children who received a D grade reacted with disbelief. There was some evidence that new friendship groups were forming around grade results. For some children, particularly in middle class schools, the knowledge that many of their friends were going to grammar school served to reinforce their feelings of inferiority.

Children tended to be extremely critical of the 11 plus across the whole of the sample. Most children were in favour of abolishing the 11 plus, but felt that some form of selection was inevitable. Secondary schools were perceived to be 'second class'. The most favoured form of selection by children was continuous assessment by the teacher over time. Overwhelmingly the children indicated that they felt their own personal opinions and experiences should be taken into account in the decision-making process about the future of the exam itself.

Authors' conclusions

The authors conclude that the children's views and opinions largely speak for themselves. However, they highlight the stress and pressure that all children experience during the preparation for the tests and the periods leading to the tests results. The groups destined for the grammar school are keen to protect their advantage, but nonetheless the majority feels the test is an unfair and unreliable mechanism for selection. The children had a number of major concerns, most keenly felt by those who perceived themselves as failures. The system encourages a process of labelling – of themselves and others. The children talked at length about the impact this had on their own sense of self-worth, whether positive or negative, as well as the social make-up of their circle of friends.
The authors go on to make a number of specific recommendations for improving the process of the test, which they argue, have some relevance to education in general. These are as follows:

Leonard M, Davey C (2001) cont'd

- Both children and parents sign the form to sit the test.
- Both children and parents receive a results letter.
- Children's results letter is written in age-appropriate language.
- Schools could bring children into school straightaway after results to minimise fear of returning to face the consequences of failure.
- Teachers should openly acknowledge the implications of the 11 plus and reassure children of their possible success in all fields of life, irrespective of grade achieved.
- Steps should be taken to reduce the stigma associated with grammar and secondary school.

A general concluding principle is that all children should feel valued by the society in which they live and that their education should contribute to positive feelings of self-worth.

Reviewers' comments

There is broad agreement with the authors' conclusions based on the evidence provided.

The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Little A (1994) Types of assessment and interest in learning: variation in the South of England in the 1980s. *Assessment in Education* 1(2): 201-222	Interest in the subject	Procedures for GCE testing at age 16	Outcome evaluation Other design	Medium

Aims

To examine school-based assessment innovations, pre-national assessment in England, and their relationship with motivation and learning

Research design

The study took advantage of the existence of different procedures for summative assessments that were in existence before the national assessment imposed uniformity of practice. Case studies of three schools, chosen because of their different school-based assessment procedures, were carried out, focusing on assessment in mathematics. The schools had adopted profiles, graded tests and numeracy profile tests. Information was collected from students by questionnaire and teachers were interviewed.

Data collection and analysis

The form of the questionnaire is not clear and how data collected from principals is not described. Mean levels of interest (across 7 questions each scored 1– 4) were compared for different schools and ability groups within them. Open-ended question responses were categorised and counted.

Author's findings

In relation to interest in maths, the mean levels of interest were highest in the school using profiles, though the differences between types of assessment were not great. Within each school ability group differences were apparent, the highest level group expressing greater levels of interest than the lower groups.
In relation to the different form of assessment, the student responses to both profiles and numeracy profile tests was more positive than to graded tests. The reasons behind the judgements varied with the type of assessment. The majority of those who were generally favourable to profiles felt that they helped identify pupil progress, understanding, strengths and weaknesses. Pupils felt that their point of view had been taken into account.

Author's conclusions

The author concludes that different types of assessment can stimulate different levels of interest expressed by students in the learning of maths, and that this influence is perceived by students to be exerted through one of at least two main ways: the first through its effect on the sense of individual control over and participation in the pace and style of learning, and the second through the specific content and quality of the test items which provide challenge, variety and interest.

Little A (1994) cont'd

'In the three cases presented in this paper, teachers were enthusiastic about the assessment process and products of their respective school-based practices. It is possible that it was the transfer of this enthusiasm to the students, rather than the intrinsic properties of the assessment procedure itself, which influenced students' attitudes to assessment, motivation and learning. The principles of ownership and control of learning by students, which lay at the heart of these innovations, also characterised the teachers' involvement in assessment, especially in the profile school. A sense of ownership and control of assessment may be as important for a teacher's sense of control of teaching as it is for a student's contrail of learning.'

Reviewers' comments

There are doubts as to the generalisability of the findings: that is, if conducted in other schools, it would not be surprising if the findings were different. Although the schools are described as case studies, there is not enough information, other than from the students, in order to form some picture of more of the impacting conditions and variables which could make the findings useful in other circumstances.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Paris SG, Lawton TA, Turner JC, Roth, JL (1991) A developmental perspective on standardised achievement testing. *Educational Researcher* 20(5): 12-20	Self-esteem attitudes to assessment	State mandated tests	Descriptive Study	Medium

Aims

To outline a developmental perspective on standardised achievement testing to determine the extent and the developmental trajectory of student frustration over standardised achievement testing to study a student perspective on standardised assessments, by asking questions such as these:
Does assessment promote motivation and pride based on personal progress? Does repeated failure on standardised tests have a cumulative impact on pupils' perceptions of their own abilities? Do students try to do their best on achievement tests? What strategies do students use when taking tests? Do test taking strategies transfer to other kinds of classroom learning? What are the developmental changes in student's attitudes and beliefs about achievement tests? What individual differences account for students different orientations to educational assessments?

Research design

Data were collected by questionnaire in three surveys. Survey one included nearly 1,000 students in grades 2 to 11, survey two included nearly 1,000 students (the same cohort) and survey three included 250 students in grades 4, 7 and 10. Questions concerned students' views about tests in general (survey 1 and 2) and about the Michigan State mandated tests in particular (survey 3).

Data collection and analysis

In survey 1, a 40-item questionnaire that was read out to students who indicated on a 5-point Likert scale whether they agreed or disagreed with the statements. In survey 2, a 20-item questionnaire about test-taking strategies was completed in the same way. In the third survey, students were asked to agree or disagree with a series of statements about the Michigan tests within weeks of taking the tests.
Methods of analysis are not stated. It is assumed that simple descriptive statistics were applied to the results from the analysis and discussion: for example, the differences that are stated between older and younger pupils.

Authors' findings

Study 1
There is a growing suspicion about the validity of test scores with increasing age.
Older students disagreed with the following three items significantly more than younger ones:

- Test scores show how intelligent you are.
- My test scores show that I am a good student.
- Test scores help to identify which teachers do the best job.

There is more disillusionment about teacher's preparation for tests. Older students indicated a higher disagreement than younger ones with the following items:

- The teacher explains to the class why we take the tests.
- The school provides useful information to my family about standardised test scores.

Paris SG, Lawton TA, Turner JC, Roth, JL (1991) cont'd

Older students report decreasing motivation to excel on standardised tests. They were less likely than younger ones to agree with the following statements:

- 'I gave my best effort when we took the test.'
- 'I want to do well on the test because my teacher really cares how well I do.'
- 'Most students try to do their best on tests.'

Older students felt less prepared to take the tests and were more likely to disagree with the following statement:

- 'I have good strategies for taking tests.'

Study 2

Contrary to expectation, students across the age range were similar in their test-taking strategies; there was no progressive and positive increase. There was no strong agreement in any grade level that the following strategies are useful: checking answers, maintaining attention, monitoring the testing time, moving away from a student who is distracting, etc.

In contrast, the authors were 'alarmed to note the significant developmental changes in the appraisals of negative strategies'. Older students were more likely to 'get tired and start filling in bubbles without reading the questions' or they just 'guess on questions that are confusing'. Adolescents are more likely to become nervous, to cheat, to have difficulty concentrating, to guess and to look for answers that matched the passage without reading the passage.

All of these strategies are designed to avoid personal effort and responsibility and thus they are detrimental to higher order thinking and intrinsic motivation.

Study 3

High achievers were more likely to report that they did well on the reading test, that the test was easy, that the test was not confusing and that they often reread parts of the passage. Persistence, appropriate strategies and positive self-perceptions distinguished the high and low achievers.

However, the results of standardised achievement tests become less valid for low achievers. Their scores are likely to be distorted by inappropriate motivation and learning strategies. To protect their self-esteem they man cease to make an effort – thus the tests .become high stakes confirmation of their lack of success,'.

Authors' conclusions

Instead of increasing motivation and test wiseness with increasing age, older students feel more resentment, anxiety cynicism and mistrust of standardised achievement tests. Older students apparently are more likely to minimise effort and respond randomly to standardised tests than are younger students. They may sabotage the tests, or use mindless strategies such as drawing patterns, etc. They may discount the tests and discredit schooling because they provide recurring evidence of low ability. These outcomes are diametrically opposed to reform aimed at promoting students higher order thinking and commitment to education.

The results of standardised achievement tests become increasingly less valid for low achievers, Their scores may be contaminated by inappropriate motivation and learning strategies in ways that affirm a self-fulfilling prophecy of low scores.

Reviewers' comments

More information about research design and methodology would have enabled a more informed response to the quality of the study.

The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	Weighting
Perry NE (1998) Young children's self-regulated learning and contexts that support it. *Journal of Educational Psychology* 90(4): 715-729	Intrinsic and extrinsic directed effort	Degree of SRL in learning experiences	Outcome evaluation Other design	Medium

Aims

To investigate relations between particular features of writing and portfolio activities and self-regulated learning (SRL) in five grade 2 and 3 classrooms in Canada.

Research design

19 grade 2 and 3 teachers were surveyed about writing and portfolio activities in their classrooms. After ranking in order of SRL, five (3 high SRL and 2 low SRL) were selected for in-depth classroom observation. Questionnaires were administered to the children in these five classes and then 10 (5 high achievers and 5 low achievers in writing) students were observed in each of the five classrooms for five sessions of writing.
The interventions were the natural differences in teaching style that are associated with encouragement of self-regulated learning: that is, the high SRL teachers offered complex activities, offered students choices, enabled them to control the amount of challenge, to collaborate with peers and to evaluate their work. The low SRL teachers were more controlling, offered few choices and their assessments of their own work were limited to mechanical features (spelling punctuation, etc).

Data collection and analysis

The key characteristic of the evaluation was to compare the actions of young children (grades 2 and 3) in classrooms that differ in respect of encouraging self-regulated learning, including the assessment practices of the teachers. Observations and questionnaire questions were directed at activities theorised to promote SRL; that is, students being given complex activities and opportunities to make choices, control challenge, collaborate with peers and evaluate their work. Teachers' interactions with students in high SRL classrooms resembled descriptions of scaffolded instruction or guided discovery.

Means and standard deviations for students' responses to the questionnaire were calculated. From these, differences in perceptions of 'control over learning' and 'support for learning' in the high SRL and low SRL classes were calculated and effect sizes calculated.

Classroom observation data were reported as anecdotal accounts of individual children.

Author's findings

There was a difference between the responses of children in high and low SRL classrooms to being asked what they would want the researcher to notice about their writing whilst looking through their portfolio. Although a large proportion of students in both contexts indicated that the mechanical aspects of writing were a focus for them (37% in high SRL and 58% in low SRL classrooms), students in high SRL classrooms alluded to the meaningful aspects (27% and 6%) and intrinsic value (21% and 3%). No student in the high SRL classrooms referred to the extrinsic value of their writing, compared with 6% of the students in the other two classrooms who indicated that they would want the researcher to see that 'most of it is right'. Students in the low SRL classrooms also were more likely to respond 'I don't know' or suggested that they didn't care.

Perry NE (1998) cont'd

Similarly, in interviews, the students observed in the high SRL classrooms indicated an approach to learning that reflected intrinsic theories of motivation. They indicated a task focus when choosing topics or collaborators for their writing and focused on what they had learned about a topic and how their writing had improved when they evaluated their writing products. In contrast, the students in the low SRL classrooms were more focused on their teacher's evaluations of their writing and how many they got right on a particular assignment. Both the high and low achievers in these classes were concerned with getting 'a good mark'.

Author's conclusions

Grade 2 and 3 students in high SRL classrooms adopted attitudes and actions that are characteristic of self regulated learners, whereas students in low SRL classrooms adopted attitudes and actions associated with defensive, even self handicapping, approaches to learning.
The author argues that the findings 'have important implications for...attending to children's motivations and designing primary classrooms environments that promote academically effective forms of SRL'.
'Researchers need to develop more valid measures of young children's motivation and SRL as well as measures that enable qualitative comparisons of students' writing.'

Reviewers' comments

The selected methods are reliably applied but they are not justified. For example, using questionnaires with 7 and 8 year olds is suspect and might be expected, as happened, to lead to uncertain outcomes. The reviewers therefore differ from the authors in the weight to be given to the findings.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	Weighting
Pollard A, Triggs P, *et al.* (2000) What pupils say: changing policy and practice in primary education. Chapters 7 and 8 in *What pupils say* by Pollard, Triggs, *et al.* London: Continuum	Attitude to tests Learning dispositions	National Curriculum Assessment and Testing	Outcome evaluation Cohort study Process evaluation	High

Aims

The project aimed to monitor the impact of the changes occurring following the passing of the 1988 Education Reform Act (ERA) in England & Wales. It focused on the impact of the changes resulting from the implementation of the ERA, particularly the National Curriculum, on teachers, headteachers and pupils.
This study addresses three key questions:

- How did pupils perceive and experience the introduction of the National Curriculum and Assessment?
- Did the introduction of the National Curriculum & Assessment facilitate or undermine the development of positive pupil learning dispositions?
- What is the significance of the recent education policy in terms of how children are understood in modern English society?

Research design

The major concern of the project was to map the educational experiences of pupils in a cohort as they developed within the new structures brought about by the ERA (the naturalistic intervention). Thus a longitudinal study of a cohort of 54 children from 9 schools was selected. In addition to this, a cohort of teachers was also identified and researched.

Data collection and analysis

Over the eight years of the study, personal interviews with headteachers, teachers and pupils were some of the most important sources of data. Self-completion questionnaires were used with teachers in order to collect demographic and career data and for attitude scales. Other procedures included observation in classrooms using both systematic quantitative procedures and qualitative approaches, with open-ended or partially structured field notes. Sociometric data on children's friendship patterns and tape recordings of teachers' interactions with children were also collected. Field notes and children's cartoon bubble completions were also used.

Data-analysis sheets were produced to record and summarise results. Sociometric diagrams were completed for each class. Systematic observation data were entered and analysed using SPSS statistical software package for the social sciences. Analyses from each of the classroom studies were compared and integrated, and entered onto the same database. This enabled comparisons across data types; for example teachers' views of children's achievement levels could be compared with pupils' perceptions of achievement levels.
Tests of statistical significance were used for the teacher interviews.

Pollard A, Triggs P, *et al.* (2000) cont'd

Authors' findings

Assessment interactions

Teachers in the early 90s tried to 'protect' pupils from the effects of the assessment reforms, which they saw as potentially damaging. Pupils in KS1 felt positive about assessment interactions. Classroom assessment was perceived as 'knowing what to do and avoiding doing it wrongly', the pleasure and pain of praise and being told off, and being told what intellectual endeavour to engage in next. Evidence from the mid 90s indicated there were fewer positive and more negative responses to questions about the assessment interaction than in the earlier years. In the final stage, the late 90s, when the pupils were in years 5 and 6, there were more summative and less implicit formative assessment tasks.

Pupils are aware of assessment only as a summative activity and use criteria of neatness, correctness, quantity and effort when commenting on their own and others' work. They drew upon the assessment discourse in the classroom for these categories. Teachers had become more accepting of a formal, structured approach to pupil assessment. There was no evidence from children that teachers were communicating any formative or diagnostic assessment to their pupils.

Pupils' judgement of their work was concerned with surface and structural features of written work, presentation, quantity and effort. Low attainers placed importance on correctness and amount, high attainers were more likely to be aware of their relative effort and living up to theirs and their teachers' expectations. Feelings of anxiety and test anxiety were reported.
A large group of children reported their home and family as accounting for their being good at something, in years 5 and 6 more pupils recognised liking a subject as being important. Few children reported liking a subject because it was easy; rather, liking it meant that they put in practice and effort. 'Enjoyment...generated the concentration, persistence, attention and willingness to practise that they saw as underpinning success' (p. 143). Explanations for not being good at something were increasingly with age given in terms of disliking and not have the necessary innate ability. Younger children gave not having enough practice as a reason whilst when they were older they indicated lack of effort. The voice of the teacher comes through clearly here.

Standardised assessment

In KS2, especially year 6, teachers were increasingly focusing on performance outcomes rather than learning processes. The pressures of standardised testing were greater in some schools than in others. Many teachers attempted to preserve pupils' self-esteem by focusing on 'doing your best' ethic and offered considerable support in practising. Children, however, seemed well aware that while trying was worthy, achieving was the required outcome.
Some children, especially high attainers in supportive schools, appeared to enjoy aspects of testing. However, others, particularly low achievers, became demotivated and dysfunctional as the difficulty of the test challenges overwhelmed them. Some 'denied' the tests and others became disruptive.
Parents, siblings and teachers mediated the official results, giving them personal meaning for children. The tests were the symbolic culmination of the children's primary education – the acid test of their achievement.
Many children associated the test with transfer to secondary school and thus to their futures and were consequently anxious about the process. Others thought that it would in any case determine the class they were put in the secondary school. Children also recognised the tests as judging what they had done and initiated not by their teachers but by the government.

Pollard A, Triggs P, *et al.* (2000)

Two-thirds of children interviewed were explicitly aware that the test results constituted some sort of official judgement of them. 'The sense that the KS2 tests were a high-stakes activity and could threaten self-esteem, social status or even lead to some form of stigma, was evidenced in many responses.' Some children, especially high attainers in supportive schools appeared to enjoy aspects of testing. Low achievers, however, became demotivated and dysfunctional as the difficulty of test challenges overwhelmed them. Some 'denied' the tests and others became disruptive.
The children's comic strips (cartoon bubbles) reflected the children's emotions, often of anger and anxiety (cf. Leonard). Their comments and drawings indicated that 'results they sought to achieve in the tests were closely associated with their sense of themselves as people and as pupils'. In KS2, especially in year 6, teachers were increasingly focusing on performance outcomes rather than learning processes. Many tried to preserve pupils' self-esteem by focusing on 'doing your best' and offered considerable support in practising tests. Children were nonetheless aware that while trying was worthy, achieving was the required outcome.

Authors' conclusions

The picture emerging from the pupils' experience of classroom assessment is consistent. Children are aware of assessment only as a summative activity and use criteria of neatness, correctness, quantity and effort in their own judgements of the quality of their work. This is drawn from the assessment discourse of the classroom. There is no evidence from the children that the teachers were communicating anything of a formative or diagnostic nature to pupils.

By the end of KS2, pupils' judgements of their work were concerned with surface and structural features of written work, presentation, quantity and effort.

The pressure of external assessment has had an impact of pupil's attitudes and perceptions. Children became less confident in their self-assessments and more likely to attribute success and failure to innate characteristics. They were less positive about assessment interactions which revealed their weakness. They reported anxiety, tension and uncertainty in relation to assessment. The assessment process was intimately associated with their developing sense of themselves as learners and as people. Children incorporated their teacher's evaluation of them into the construction of their identity as learners.

Teachers increasingly focused on performance outcomes rather than learning processes. Low achievers particularly became demotivated and overwhelmed by assessments. The consequence of an increased focus on assessment was increased differentiation. The Y6 tests were the symbolic culmination of children's primary education and the acid test of their achievement.

The anxiety children felt was arguable a consequence of the sense that they were exposed to greater risk as their teacher raised the stakes. Relating this to Bernsteins's association of evaluation with power, the authors concluded that the children felt the power of their teachers as assessors, especially as the distributors of rewards and punishments, including the giving and withholding of approval. 'It is clear that for these children assessment had more to do with pronouncing on their attainments than with progressing their learning.'

Reviewers' comments

There is broad agreement between reviewers and authors. More quantification of the themes emerging in the data would strengthen the findings and help interpretation. There is no discussion of alternative interpretations of the data, such as the possibility that the changes were associated with age and maturity rather than outside factors.

What I feel and think about myself in relation to learning				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Reay D, Wiliam D (1999) 'I'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal* 25(3): 343-354	Self-esteem Attitude to tests Tests anxiety Self perception as learners	National Curriculum Tests	Process Evaluation	High
Aims				
To explore the extent to which year 6 children's perceptions of the (national curriculum) tests contribute to their understandings of themselves as learners				
Research design				
A small-scale study conducted in one classroom of a London primary school. The report focuses on data gathered in on year 6 class over the Easter term (i.e. January to April) 1998. During this time, the children were being prepared for the testsand the researcher spent over 60 hours observing teaching and learning processes in the classroom. They also amassed extensive field notes documenting both changing pedagogic approaches and the children's responses to them. All the students were interviewed in focus groups and half the class were interviewed individually about their attitudes towards, and feeling about, impending National Curriculum Tests.				
Data collection and analysis				
Findings are presented in terms of verbatim quotations from focus groups and interviews and vignettes of particular classroom events that were observed. For example, 'In March 1998 the children were working their way individually through an old science test paper. Fumi protested at the beginning of the session when told the children were expected to work on their own, telling the teacher, 'But we're used to working together.' Every few minutes she would sigh audibly until eventually the teacher came across to where she was sitting and proceeded to put lines through a number of test questions, commenting 'Don't try and do these. They'll be too difficult for you. Answer the easy ones'. Fumi struggled on for a few more minutes. It was clear to the researcher and the children sitting near her that she was crying. After a few more minutes she got to her feet, pushing her chair out of the way and stormed out of the classroom, sobbing. 'He thinks I'm thick. He thinks I'm thick. He wants all the other to think I'm thick'.'				
Authors' findings				
The findings are the points illustrated by quotations and observations: for example, after quoting from Hannah, 'For Hannah what constitutes success is correct spelling and knowing your times table. She is an accomplished writer, a gifted dancer and artist and good at problem solving yet none of those skills make her a somebody in her own eyes. Instead she constructs herself as a failure, an academic non person, by a metonymic shift in which she comes to see herself entirely in terms of the level to which her performance in the SATs is ascribed.' There is a description of the class as being at 'fever pitch' because of the impending tests and the teacher's own anxieties were evidence in the way he berated the children for poor performance in the practice tests.				

Reay D, Wiliam D (1999) cont'd
All the children, except the brightest boy, expressed varying degrees of anxiety about failure, with girls expressing more anxiety than boys did. They took the tests seriously and wanted to do well. Children were reported as expressing concern about the narrow focus of the tests and not being able to produce their best under strict and unfamiliar test conditions. The children recognised that the tests were about how well they have been taught but still had 'a sense of unease...about what the SATs might reveal about themselves. Some of the children seemed to be indicating far reaching consequences in which good SATs results were linked to positive life prospects and, concomitantly, poor results mean future failures and hardships'. The children were reported as equating cleverness with doing well in the tests. They were beginning to view themselves and others differently, in terms of test results. 'As the term progressed, children increasingly referred to the levels they expected themselves and others to achieve'. Their talk raised concerns about the crudeness of the assessment to which pupils have access.
Authors' conclusions
The main conclusions are as follows: 1. As time went on and the tests became closer, changes occurred in the content and methods of teaching which in turn had an impact on the relations among peer groups of children. 2. The students studied are well aware of the effects of NC assessment. 3. Threats to schools posed by poor test results puts teachers under pressure to increase scores 'irrespective of the consequence for students' achievement in wider terms' (emphasised by the authors). 4. The narrowing of the focus of the assessment and emphasis on achieving the highest scores possible 'produces a situation in which unjustifiable educational practices are not only possible but also encouraged'. 5. Such practices rob the National Curriculum assessment of the power to say anything useful about what the students have learnt. 'The more specific the Government is about what it is that schools are to achieve, the more likely it is to get it, but the less likely it is to mean anything.'
Reviewers' comments
No differences in terms of the data provided. The findings are well illustrated in the selected verbatim quotes and the classroom observations. Although a good deal of direct evidence is provided, it is not possible to be sure that this reflects the picture that the full data convey and does not preclude other interpretations

The energy I have for the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Roderick M & Engel M (2001) The grasshopper and the ant: motivational responses of low achieving pupils to high stakes testing. *Educational Evaluation and Policy Analysis* 23(3): 197-228	Effort	Threat of retention as result of tests	Process evaluation	Medium

Aims

This study investigates the impact of the threat of grade retention on 102 low-achieving students who face grade retention if they do not improve their test scores on the Iowa Test of Basic Skills (ITBS). This study examines the pre-testing experiences of 102 low-achieving sixth- and eighth- grade students in five schools who faced the Chicago Public School's (CPS) promotional test score cut-offs in 1999. It draws on interviews conducted with students before testing, assessments from their teachers, and their school records to examine three central questions. First, is there evidence that the CPS policy leads low-achieving students to work harder and has an impact on their learning goals? Second, how do teachers and families shape that experience for students? Third, is there a relationship between students' responses to the policy and their learning gains and promotional outcomes? In essence, does hard work pay off?

Research design

In 1996, the Chicago Public Schools (CPS) introduced an initiative meant to end social promotion. CPS students in the third, sixth and eighth grades must achieve a minimum score on a standardised test in reading and mathematics, the Iowa Tests of Basic Skills (ITBS) in order to be promoted to the next grade. Schools are provided with funds to extend instructional time for students who are deemed at risk of failing through the Lighthouse after-school program. Students who do not meet the promotional criteria are required to participate in a special summer school program, Summer Bridge. Those who fail again are retained in their grade or, if they are 15, attend alternative schools called transition centres. In the first two years of the policy, CPS retained 20% of eligible third graders and approximately 10% of e sixth- and eighth-grade students.

The sample consisted of schools with the highest concentration of students at risk of failing to meet the cut-offs and, within these schools, students who were likely to face summer school or retention or both. Approximately a quarter of the sample was at high risk and 57% at moderate risk.
The intervention was designed to address the problem of the lack of achievement norms with an initiative meant to end social promotion. The aim was to improve achievement by threatening grade retention which carries not only academic but also social consequences for the students.

Data collection and analysis

Answers to the four sets of questions were coded to identify primary themes in students' description of their experiences in the year before the test, their work effort in school and their approach to test preparation. Students were then grouped into four categories that represented the common themes.

Authors' findings

The first group (53%) were those who were working harder in school as a result of the intervention. They perceived the policy had altered their experiences in school and attitudes towards learning and led them to increase their effort. They reported greater attention to class work, increased academic press (high expectations) from teachers, greater academic effort in and out of class. A higher proportion of these children came from low and moderate risk of retention groups.

Roderick M & Engel M (2001) cont'd

The second group (9%) were those working harder but outside school, supported by other adults. Most of these students had supportive parents. They were evenly spread across gender and grades and race.
The third group was the 'worrying but not working' group, comprising 34%. These students seldom related what they were doing in school to preparing for the ITBS. There was a higher proportion of 6th graders, males and Latinos in this group and a high proportion of students in the school that was lowest in group 1 students.
The fourth group comprised four students (4%) who were the most highly skilled in the sample and had already met targets in at least one subject.
Across the groups, there were differences in age, gender and race. 8th graders worked harder than 6th graders, males less than females and Latinos more likely to be worrying and not working than Afro-Americans. There is evidence of a school effect in that teacher support in relation to the policy is reflected in the outcomes. Schools in the same district produced different outcomes and different levels of teacher support. A high teacher support school is associated with more effort. This included creating an environment of social and educational support, working hard to increase students' sense of self efficacy, focusing on task centred goals, making goals explicit, using assessment to help pupils succeed and creating cognitive maps. They also adopted a strong sense of responsibility for their students.
Low teacher support school in a second school in the same district, included teachers not seeing the target grades as attainable, not translating the need to work harder into meaningful activities, not displaying recognition of change and motivation on the part of students, not making personal connections with students in relation to learning goals.
Actual test results were collected for these students and new groups were created to compare effort with outcome. Groups 1 and 3 were combined. Those not making an effort were divided according to personal and home problems: those not making an effort with problems and those without. 21% of the effort students were retaking and 20% passed at first or second attempt. In contrast, 64% of students in the no effort group no problem group were retained, of whom 36% passed. 42% of the no effort and problems group were retained.

Authors' conclusions

The authors conclude that low-achieving students will always react negatively to policies that place a strong emphasis on achievement. The majority of students in the sample responded positively to the policy. The need to reach the test score cut-offs became a factor that shaped their attitudes toward school and essentially transformed the value that they placed on learning, at least in the short term. Their responses suggest that creating incentives for low-achieving students through goals that provide an opportunity for feedback, a tangible reward and a way to construct meaning regarding learning may have a positive impact on their motivation and effort in school.

Students with the lowest skills were the least likely to respond positively. This suggests that, even if being promoted was something the students valued, they might not have felt that the goal was attainable or that they could influence their own outcomes, given their low skills or lack of support, or both. Students with low motivation were more likely to lack external support and to have problems outside of school that created barriers to their engagement in their schoolwork.

Roderick M & Engel M (2001) cont'd

Policies that rely heavily on student motivation to improve achievement may place students with the lowest skills in a very difficult position. These students face the greatest task and at the same time often have the fewest resources to accomplish that task These students are also less likely to be able to translate their desire to be promoted into substantial work effort.

The degree to which students respond to high-stakes testing through motivation and increased work effort is an important predictor of their learning outcomes.

Teachers play an important role in shaping student's outcomes in high-stakes testing environments. Helping students understand the policy, making them feel supported and efficacious in achieving goals and structuring meaningful activities are all essential components.

High-stakes testing, using the negative incentive, means that some students will fail. This makes 'sacrificial lambs' of the most vulnerable. Teachers need to address their diagnostic capacity and support to address chronic learning, health, familial and other multi-dimensional problems.

Reviewers' comments

These findings relate to low-achieving children, with relatively low skill gaps to close. It is not clear whether it is the threat of retention or the high stakes assessments that increased effort.
We do not know what happens with higher cut-off scores when more students might, like the lowest achievers here, find the gap too great for them to bridge with their resources.

How I perceive my capacity to undertake the task				
Study name	**Outcomes reported (relevant to the review)**	**Intervention**	**Study type**	**Weighting**
Schunk D (1996) Goal and self-evaluation influences during children's cognitive skill learning. *American Educational Research Journal* 33(2): 359-382	Self-regulatory processes Self-efficacy	Goal orientation and self-evaluation conditions	Outcome evaluation RCT	High

Aims

Two studies together had the purpose of 'exploring the operation of self-regulatory processes among children during cognitive skill learning' in mathematics.
Study 1 hypotheses that self-evaluations of capabilities would positively affect motivation, self-efficacy, learning goal orientations and skills. Combining learning goals with self-evaluations would prove most effective. (Learning goal refers to what knowledge and which skills students are to acquire; a performance goal denotes what task students have to complete).
Study 2 was designed to test the prediction that learning goals would lead to higher self-evaluation scores and achievement outcomes than performance goals.

Research design

Study 1: 44 4th grade students (18 girls and 26 boys) from one school
Study 2: 40 4th grade students (20 boys and 20 girls) (different from those in Study 1) from one school
Students were randomly assigned 'within gender, ethnic background and classroom, to one of four experimental conditions: learning goals with self-evaluation, learning goals without self-evaluation, performance goals with and without self-evaluation. Students received 45 minutes instructional sessions over 7 days. Children assigned to the same condition met in small groups with one or two female teachers from outside the school.' Each teacher worked with all four experimental conditions.

The instructional packages covered six major types of fraction skills and the seventh was a review package. 'In each package, the first page explained the relevant operations and exemplified their application. Each of the following pages contained several similar problems to be solved using the depicted steps. Each set included more problems than children could complete during the session.'

At the start of the session the teacher gave the goal instruction appropriate for the children's condition, then verbally explained and demonstrated the relevant fraction operations. 'After this modelled demonstration phase (about 10 minutes) students engaged in a hands-on activity with manipulatives and cut outs and solved a few practice problems (guided practice). Once the teacher was satisfied that children understood what to do, children solved problems alone during independent practice for the remainder of the session (about 25 mins.).'
The goals were presented as follows: for the learning goal groups the teacher said 'While you are working it helps to keep in mind what you're trying to do'. She stressed the goals of learning to solve problems: 'You'll be trying to learn how to solve fraction problems where the denominators are the same and you have to add the numerators.' These instructions were varied according to the type of problem being tackled in the session.

For the performance goals groups the teacher gave the same initial instruction but the session specific goal made no explicit mention of learning.
(The author recognised that the difference was small but 'to ensure that the conditions were distinguished and that the children understood their instructions, the teacher verbalised the instructions at the start of each session'. In addition, the teacher asked children to repeat the instructions and after this asked if that sounded reasonable.)
Self-evaluation: the children (in the relevant groups) judged their fraction capabilities at the end of each

Schunk D (1996) cont'd

of the first six sessions. 'The materials and procedure were identical to those of the retest self-efficacy assessment, except that that children judged how certain they were they could solve the types of fraction problems covered during that session.' (In the self-efficacy assessment they were shown briefly pairs of problems for about 2 seconds which allowed assessment of problem difficulty but not actual solution.)
Children in the no self-evaluation group did not engage in end-of-session evaluation but completed an attitude questionnaire.
Study 2: As for study 1 except that all students received the opportunity for self-evaluation but this was conducted once, at the end of the program, rather than six times. (The purpose was to reduce the overwhelming influence of self-evaluation found in Study 1 so that the effect of the goal orientation could be investigated.)

Data collection and analysis

Instruments used as pre- and post-test self-completion measures were goal orientation inventory, as basis for measurement of goal orientation; a self-efficacy scale based on earlier work by author; a skill test relating to the mathematics being learned and created in two parallel forms, for use as pre- and post-test. The problems in the skill test were administered one at a time and the time spent on each was taken as a measure of persistence.
ANOVA was used to investigate any significant between condition differences on pre-test. The three post-test measures were analysed with MANCOVA with goal orientation as the experimental factor and the corresponding pre-test as covariates.
In Study 1, product moment correlations were computed among lesson performance (number of problems completed) and post-test measures (goal orientations, self efficacy, skill, persistence).
In Study 2, correlational analysis was used to explore relations between instructional session measures (number of problems completed), self-efficacy for learning, self-evaluation, self-satisfaction and goal perceptions.

Author's findings

Relevant finding for this review are those relating to goal orientation, on the assumption that summative assessment is related to performance goal and to self-evaluation.
In Study 1, the effect of goal orientation was apparent only when self-evaluation was absent. Children under self-evaluation conditions and under learning goals with no self-evaluation solved significantly more problems that did those with performance goals and no self-evaluation. Self-evaluation scores for performance goals and for learning goal were not significantly different.
Study 2 (where all students engaged in self-evaluation) showed significant effects for self-efficacy and for skill. The learning goals group scored higher that the performance group condition on both measures. In relation to goal orientation, the performance goal group reported higher ego orientation and work avoidance orientation than did the learning goals group.
The results of these two studies show that providing students with a goal of learning to solve problems enhances their self-efficacy, skill, motivation and task goal orientation; and that these achievement outcomes are also promoted by allowing students to evaluate their performance capabilities or progress in skill acquisition.
The results of the studies differ in that Study 2, but not Study 1, supports the hypothesis that 'combining learning goal with self-evaluation raises achievement outcomes more than does combining a performance goal with self evaluation... A daily assessment of capabilities is clearly intensive and should communicate to children that they are becoming more skilful. When self-evaluation is so salient, the type of goal may make little difference. In contrast, the single self-assessment in Study 2 may not have made it clear that subject had become more competent'.

Schunk D (1996) cont'd
Author's conclusions
The author provides a theoretical explanation for these findings as follows. Emphasising to students that their goal is to learn to solve problems can raise their self-efficacy for learning and motivate them to regulate their task performance and work diligently. Self-efficacy is substantiated as they observe their progress in skill acquisition. Higher self-efficacy helps to sustain motivation and skilful performance. The study results 'support the idea that self-efficacy is merely a reflection of prior performance'. The results suggests that treatment conditions differed in the extent they conveyed a sense of learning progress to students, which enhanced their self-efficacy, self-regulatory activities and learning.
Reviewers' comments
A significant conclusion is the following: 'Among children who are capable of evaluating their capabilities, self-evaluation may be a useful adjunct to testing as a means of assessing students' skills and or providing information to use in designing instruction. Although learning goals and self-evaluation are not necessary for all classroom activities, the present results suggest that, when combined with a sound instructional program, they facilitate self-regulated learning and achievement out comes'.

APPENDIX D: Conference report

Report of the consultation conference held on 5 March 2002 to consider the outcomes of the systematic review of research on the Impact of summative Assessment and Tests on Students' Motivation for Learning

Introduction

The conference was organized by The Assessment Reform Group (ARG) and the Centre for Assessment Studies (CAS) of the Graduate School of Education, University of Bristol. It was attended by 45 invited educational professionals: seven local authority or independent advisers, 11 policymakers from government or government agencies throughout the UK, four teachers (more had been invited, but were unable to leave their schools on the day), eight teacher educators, six academics with research interests in assessment and nine with research interests in educational policy. The purpose of the conference was to discuss the findings and implications of a systematic review of research in relation to the impact of testing on students' motivation for learning. The review had been carried out by Wynne Harlen (Project Director) and Ruth Deakin Crick with the assistance of the Assessment and Learning Research Synthesis Group (ALRSG), a project group drawn together by ARG.

The review was funded by the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) and by the Nuffield Foundation. The conference took as its focus, the final stage of the review, the synthesis, and was designed to engage policymakers, research users, researchers and the review team in an examination of the review's findings and implications.

Overview of the conference programme

The conference was designed to enable as much dialogue as possible with plenary presentations made by Wynne Harlen and Ruth Deakin Crick being used to initiate the discussions of small breakout groups. The morning and afternoon sessions each took a different theme and are reported separately in the sections that follow.

The plenary presentations for the morning session, chaired by Gordon Stobart, related to the methodology of the review and the findings. Five breakout groups were then asked to consider: *How does this evidence resonate with your experience?* with a focus on any surprises, emerging questions or additions that they might wish to make. The session concluded with lunch and a summary of the discussions (captured on flipchart sheets) was provided by Richard Daugherty as feedback at the

beginning of the afternoon session. The focus of the rest of the afternoon session was the implications for policy and practice of the review and this was introduced by the review authors with an overview of the findings relating to possible action to be taken in order to reduce the negative impact and increase the value of summative assessment.

Participants were then formed into five groups according to their professional interest to address the question: *What are the implications, from your perspective, for policy and practice?* Feedback on this occasion was by brief summaries from the group rapporteurs to the plenary group, with the detailed flipchart sheets being retained for transcription and summarizing later.

The closing input to the conference was then offered by Carolyn Hutchinson, Head of Assessment Branch, Scottish Executive Education Department. She reflected on the day's proceedings and outlined how *Assessment for Learning* was being addressed in Scotland's Assessment Development Programme with details set out in a handout (carolyn.hutchinson@scotland.gsi.gov.uk).

Session 1 (morning)

Wynne Harlen presented the background to the review, outlining the reasons for undertaking it and the roles of the various people in facilitating it (the EPPI -Centre team, the Nuffield Foundation, the ALRSG and ARG itself). She charted the main features of the literature, beginning with the positions taken on the impacts of summative and formative assessment on standards of performance in schools. Establishing summative testing as the focus, she then outlined the main tenets of current theories linking motivation and learning, briefly reviewing a broad range of concepts such as goal theory, attribution theory, intrinsic and extrinsic motivation. Ruth Deakin Crick then described the systematic review methodology, beginning with the identification of the review question:

- What is the evidence of the impact of summative assessment and testing on students' motivation for learning?

Ruth set out the systematic nature of the EPPI review process. The initial step, a wide-ranging search for relevant papers and reports, led to the identification of 183 studies. Subsequent steps applied increasingly specific inclusion and exclusion criteria (from keywording at the top level to in-depth critical reading criteria in the final stages), which ultimately whittled the body of evidence down to 19 pieces of work. The clear delineation of 'summative assessment' and 'motivation for learning', as the independent and dependent variables, was crucial to the inclusion/exclusion process. This meant that studies focusing on the impact of motivation (disaffection, incentives, etc.) on performance, as measured by summative assessment, were excluded since in such cases the independent and dependent variables were interchanged.

Following this, Wynne reviewed the main findings, grouping studies according to the aspects of motivation they reported.

Feedback from the morning discussions

How does this evidence resonate with your experience?

The findings of the review were viewed as being broadly in sympathy with professional experience. In general, however, the feeling was that the low number of relevant studies was surprising, given the intense interest in testing, particularly in the UK and US.

It is in the nature of such complex issues that the discussions raised many interesting questions and suggestions, many of which could not be accommodated in the current review owing to its highly structured and specific nature. The various observations were transcribed from the flipcharts drawn up by the five groups and are summarized in point form below. Clearly no two groups took exactly the same approach in addressing the breakout agenda, but the following perceived needs were distilled from all of the responses, i.e. the need:

- for definitions of high and low achievement when considering differential impact
- to consider the impact of peer summative assessment on motivation for learning
- for more empirical work on learning style preferences
- to distinguish between assessment of individuals and measurement of cohorts, when considering impact
- to disseminate diagnostic information on summative tests, which it was felt was held by examination bodies, to schools to enable the review of individual and cohort impact within the schools
- to recognize that current testing is changing the way teaching is carried out and to assess how this impacts on motivation for learning
- to assess the cost benefits of taking the evidence of the review on board
- to consider how to achieve complementarity between formative and summative assessment
- to disaggregate the data gender/ethnicity/learning disposition, etc.
- to consider the impact of contextual factors, such as school/parent/teacher values and anxieties

Session 2 (afternoon)

Session 2 began with Wynne Harlen reporting some further analyses which addressed the issue of differential impact of summative assessment in respect of gender, age and achievement level. She also reported initial findings with regard to the conditions of testing, the effect of high stakes and the impact on teaching, all of which mediate the impact of tests on students motivation. She also offered a starting point for some possible ways forward, giving as examples the suggestion

made by Kohn (2000), in his book *The case against standardized testing*[1]. Wynne outlined a series of proposals for 'what needs to be done' including:

- stopping drill and practice approaches to tests
- reducing the impact of testing on teaching
- teachers emphasizing learning goals in preference to performance goals
- introducing or emphasizing forms of assessment other than tests
- providing information to students about the purposes, meanings and requirements of assessments
- developing a positive assessment ethos in schools
- improving and restructuring tests, keeping them to a minimum
- reducing the high stakes context for tests, for example the pressure of league tables on schools in some of the UK countries
- ending the use of tests for multiple purposes including teaching quality evaluation

Wynne also referred to the findings of the OECD/PISA[2] project which lent support to the importance of attention to teaching in ways that give students some control over their learning. At the conclusion of Wynne's presentation, Ruth Deakin Crick juxtaposed two sets of factors that may be considered to hinder or promote learner-centred practice in teaching. She then posed the second session's working question:*What are the implications of the findings, from your perspective, on policy and practice?*

To discuss this question, participants were grouped according to their role in education;: teachers, advisers, government and government agency policymakers, assessment researchers and policy researchers.

Feedback from afternoon discussions

The policymakers' group considered that there was a serious message in the report for policymakers and that the focus of this message is value for money. The prospect of administrative 'melt-down' arising from system characterized by over-testing, the need to address the gap between high and low achievers, the need to provide more appropriately for gifted and talented students and the need to examine the compatibility of the current assessment system with 21st century learning and skills, were all considered to indicate inefficient use of learning time and public funds. In formulating the 'message' for policymakers, clear and concise language was advised, for example in distinguishing between the implications set out for policy versus practice.

In schools, there is a need to ensure that senior managers understand the research and its implication in order that they can best set priorities and enable teachers to 'take risks' and change what might be long-

[1] Kohn A (2000) *The case against standardized testing, raising the score, ruining the schools*. Portsmouth, New Hampshire: Heinemann
[2] OECD (2001) *Knowledge and skills for life: first results from the PISA Project.* Paris: OECD

established practice. The national numeracy and literacy strategies were considered to have responded to the worst excesses of drill and practice but the National College for School Leadership was judged to have a major role in developing senior managers' level of understanding of assessment beyond management information and accountability needs.

The suggestion was made that LEA appointments supporting the Key Stage 3 National Strategy in England might be an appropriate vector for promoting assessment for learning. The promotion of a wider awareness, in particular among policymakers, of the 'fuzziness' of assessment and the confidence limits in grading and marking bands was also felt to be important targets for follow-up action. Such imprecision undermines the basis of much accountability-related legislation and policy, and this review may help to bring home the message of the negative impacts of such assessments.

Continuing the theme of promoting the review's implications with policymakers, it was felt that the following aspects should be stressed:

- current inefficient use of public money
- unsustainable reliance on unreliable data
- focus on raising standards (as appropriate assessment for learning is known to achieve)
- waste of learning time on practice tests and marking
- closing the gap between higher and lower achievers

Finally, it was thought timely for policy to pre-empt the potential collapse of the current testing/examination regime.

The advisers' group highlighted the need to generate respect for the wider range of achievements available to lower attaining students. Such students are believed to be more susceptible to the negative impacts of testing on their motivation for learning than perhaps their higher attaining counterparts. The need to review how assessment is applied at different stages of students' development was also highlighted.

The polarization of formative and summative assessment was considered to need addressing and teachers' competence in their use to be improved to ensure a positive impact on students' motivation for learning. The reduction of teachers' anxiety in relation to assessment was felt to be an important part of this process. While it was thought that the use of national assessments would remain as an element of the accountability of schools and teaching, it was also acknowledged that country specific differences would create fundamentally different policy drivers. Finally, it was felt important to engage ICT solutions appropriately and efficiently in support of teachers and assessment.

The teachers' group echoed a number of earlier points by reinforcing the importance of teachers' professional development and the need to reduce the high stakes associated with much of today's summative assessments. They considered that there should be continuous access

to assessment levels and that 'calendar-based' testing should be resisted. They also revisited the need to minimize didactic teaching.

The policy researchers' group developed a concept map (Figure 1) to assist their analysis and discussion. They felt that work was needed to open up policy concerns by providing appropriate case histories and that influential networks should be mobilized to ensure that motivation to learn becomes a focus of current debates. It was considered necessary to have a clear understanding of how policy works and 'where the policy makers are coming from'. Assessment should be treated as a pedagogic issue for policy change with teacher research networks being supported and incentives being offered to schools. As with an earlier group, the role of technology was considered as a future area for expansion and the advice given was that the system should work smarter, not harder.

The fifth group, of assessment researchers, focused on research needs and argued for implementation studies of different assessment models to analyse more fully the impact on motivation for learning. For example, little is known of the potential effect of students being involved more in the process of testing through self-assessment and the interpretation of grades, marks etc. In relation to the review's findings, it was felt that aspects of applicability and generalizability (transferability) would need attention as several of the papers providing evidence did so from contexts considerably different to those found in the UK. Several observations were offered:

- Research into the relationship between motivation (the independent variable) and performance outcomes (the dependent variable) was considered another rich source of evidence which needed to be tapped in a similarly thorough fashion in order to understand this complex issue more fully
- The review may be a viable means of causing policymakers to consider change but further research or the incorporation of excluded research may be needed to guide the necessary changes.
- Notwithstanding current policy and the legislative framework, much can be done by teachers and schools to learn from the positive aspects of the study to change practice.

Finally, the group posed the question: *What would be an optimum design for the balance between internal (to schools) and external assessment?*

Figure 1: The fourth group's concept map