Assessment in Education, Vol. 10, No. 2, July 2003

Carfax Publishing
Taylor & Francis Group

Testing and Motivation for Learning

WYNNE HARLEN & RUTH DEAKIN CRICK
Graduate School of Education, University of Bristol, 35 Berkeley Square, Bristol BS8 1JA, UK

ABSTRACT *This paper presents the procedures and findings of a systematic review of research on the impact of testing on students' motivation for learning. The review was undertaken to provide evidence in relation to claims that, on the one hand, testing raises standards and, on the other, that testing, particularly in high stakes contexts, has a negative impact on motivation for learning that militates against preparation for lifelong learning. Motivation is considered as a complex concept, closely aligned with 'the will to learn', and encompassing self-esteem, self-efficacy, effort, self-regulation, locus of control and goal orientation. The paper describes the systematic methodology of the review and sets out the evidence base for the findings, which serve to substantiate the concern about the impact of summative assessment on motivation for learning. Implications for policy and practice are drawn from the findings.*

Introduction

In this paper we report a review of research carried out to identify evidence of any impact of testing and other forms of summative assessment on students' motivation for learning. Our findings are framed by the reasons for the review, its funding, timing, methods and focus and the meaning of key terms; thus discussion of these things forms an important part of this paper. The review was conducted during 2000 and 2001 following the procedures for systematic review of research in education being developed at that time by the government funded Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre). These procedures differ in several respects from those of narrative reviews. We therefore begin by setting out the background to the review, our view of the meaning of key terms and an account of the review methodology. The main section gives the findings of the review. We conclude with some implications for policy and practice that emerged from discussing the findings with policy makers and practitioners.

Background

There were two sets of circumstances coinciding to bring about the particular focus of this review: one relating to the topic and the other to the review methodology.

ISSN 0969-594X print; ISSN 1465-329X online/03/020169-39 ©2003 Taylor & Francis Ltd
DOI: 10.1080/0969594032000121270

These circumstances help to explain the choice of what was included and what was not covered by the review.

The Growth of Testing

The need for a review of the impact of testing on motivation for learning was identified as a result of events following the review of research on classroom assessment by Black and Wiliam (1998). Their review revealed strong evidence that improving formative assessment can significantly raise standards of attainment. However there was concern, based on the growing international research evidence, particularly from the USA and UK, where assessment for summative purposes has burgeoned in the past decade, that the use of tests not only inhibits the practice of formative assessment but has a negative impact on motivation for learning. Moreover the evidence suggested that the effect was greater for the less successful pupils and thus tends to widen the gap between higher and lower achieving pupils.

The association of testing with a negative impact on motivation contrasts with the view, widely held among politicians, parents and some of the education community, that testing pupils raises standards. Kellaghan *et al.* (1996) identified six propositions put forward in favour of this view. 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 USA, 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. They also reflect the view that testing raises standards; a view that appears to be supported by increases in test scores following the introduction of tests. Research into testing programmes, however, has been used to show that increase in test scores over time is likely to be due to greater familiarity of teachers and pupils with the tests rather than increasing learning (eg Kohn, 2000; Koretz, 1988, 1991; Linn, 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 & McClune, 2000). In these circumstances teachers make little use of assessment formatively, to help the learning process (Broadfoot & Pollard, 2000; Osborn *et al.*, 2000; Pollard *et al.*, 2000). In other words, high stakes summative assessment squeezes out formative assessment.

In the USA the growth of external tests has been charted by Clarke *et al.* (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. In England, too, there is test-inflation. A survey by the Qualifications and Curriculum Authority conducted in 2000 found that the introduction of national tests brought with it an increase, not a decrease, in use of other tests. It is estimated that the average student in England

takes 60 tests between the ages of 4 and 18 (Professional Association of Teachers, 2000). The USA and England now vie for the title of 'most tested nation'. When Resnick and Nolan (1995) claimed this title for the USA, noting that there were few countries today that gave these formal examinations to students before the age of 16, they were not taking account of the rapid, and what may have seemed untypical, changes in the UK. However the USA remains the country where 'short-answer questions and computational exercises presented in formats that can be scored quickly and 'objectively' is the typical style of testing' (Schoen *et al.*, 1999, p. 446).

It is not only external tests that impact on pupils. Research (Black, 1993; Crooks, 1988; Pollard *et al.*, 2000) shows that, in practice, teachers' assessment has more of the characteristics of summative than formative assessment and often emulates external tests in the assumption that this represents good assessment practice. 'The evidence is that with such practices the effect of feedback is to teach the weaker pupils that they lack ability, so that they are de-motivated and lose confidence in their own capacity to learn' (Black & Wiliam, 1998, p. 18).

As a result of the explosion in testing, it has become for most students in England, most of the USA and in many other 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, p. 46), but rather a frequent experience which can have an undesirable impact on motivation for learning. Thus this review includes classroom tests and assessment that have summative purposes, as well as external tests. It excludes classroom assessment with a formative intent.

Earlier Reviews of Testing and Motivation

Reviews of research relating to testing have typically covered a range of impacts on students, teachers and the curriculum. Of those giving specific attention to testing and motivation, the work of Kellaghan *et al.* (1996) is the most relevant. 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. The review of Deci and Ryan (1985) also provides research evidence that assessment of the kind that takes away control from the learners reduces intrinsic motivation and leads to 'surface' learning.

Crooks (1988) 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 a 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 outside school. Crooks also drew attention to research that indicated problems associated with extrinsic motivation in tending to lead to 'shallow' rather than 'deep' learning.

Ames' (1992) review was concerned to look at achievement goals and to identify the situations and instructional strategies that lead to motivation towards desired goals. She contrasted learning goals with performance 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 have 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 increases 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 relationship between test anxiety and test performance. Although there were also studies which reported no relationship, McDonald concluded 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.

Two reviews, by Madaus and Clarke (1999) and McNeil and Valenzuela (1998) were presented at a conference on High Stakes Testing K–12 held at Harvard University in December 1998. They had a specific focus on research relating to issues of high stakes testing in the USA. Madaus and Clarke focused on the impact of high stakes testing on minority students, drawing mainly on research conducted at Boston College's Centre for the Study of Testing, Evaluation and Educational Policy. They used the research to identify not only 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 conclude that:

- 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 programmes have been shown to increase high school dropout rates—particularly among minority student populations. (Madaus & Clarke, 1999, p. 1)

McNeil and Valenzuela (1998) 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 schools where administrator pay is tied to test scores and where test scores have been historically low. (McNeil & Valenzuela, 1998, 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' (McNeil & Valenzuela, 1998, 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.

The Meaning and Importance of Motivation for Learning

The complexities of life in the twenty-first century have brought to the forefront of educational thinking the need for students in schools to be supported in developing the capabilities, qualities and dispositions for effective lifelong learning. This adds to the importance of embracing motivation for learning as a goal of education at all levels. It also means that if, as suggested, some assessment practices are reducing motivation for learning, this is clearly of concern. However, motivation is not a single or a simple concept and so it is necessary to consider the range of factors which constitute motivation for learning, and the kind of motivation that is needed for learning how to learn and for lifelong learning.

Motivation for learning is a complex overarching concept, which is influenced by a range of psychosocial factors both internal to the learner and present in the learner's social and natural environment. The American Psychological Association's (1997) *Learner Centred Principles* focus on factors that are internal to, and under the control of the learner, as well as taking account of the environmental and contextual factors which interact with those internal factors. Of their fourteen principles, three deal directly with motivation for learning. The first of these has to do with the motivational and emotional influences on learning, which are affected by the

learner's emotional state, beliefs, interests, goals and habits of thinking. The second refers to the learner's creativity, higher order thinking and natural curiosity that contribute to intrinsic motivation to learn. Intrinsic motivation for learning is stimulated by tasks of optimal novelty and difficulty, relevant to personal interests and providing for personal choice and control. The third principle has to do with the effect of motivation on extended learner effort and guided practice—without motivation to learn, the willingness to exert this effort is unlikely without coercion.

These three broad principles indicate the range of factors that have to be taken into account when considering motivation for learning. They have to do with the learner's sense of self, expressed through values and attitudes; with the learner's engagement with learning, including their sense of control and efficacy; and with the learner's willingness to exert effort to achieve a learning goal.

Learners' Sense of Self

In describing the key determinants of motivation for learning, McCombs and Whisler (1997) identify self-awareness and beliefs about personal control, competence and ability, clarity and salience of personal values, interests and goals, personal expectations for success or failure and affect, emotion and general states of mind as central factors. These relate to the notion of a 'learning identity'—those beliefs, values and attitudes, which the learner holds about and towards themselves and which have an influence on their goal orientation—and to their sense of efficacy as a learner.

A person's perceptions of the causes of success and failure are of central importance in the development of motivation for 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.

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 learning, 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 (also known as task involved or mastery 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 (achievement or ego-involving) goals (Ames, 1990a,b; 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.

Intrinsic and Extrinsic Motivation

Educational psychologists and researchers distinguish between intrinsic and extrinsic motivation. Intrinsic motivation, meaning that learners find interest and satisfaction in what they learn and in the learning process itself, 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. 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. There is a considerable body of opinion and evidence that suggests those 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).

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 what is required to obtain the reward. However, opinions differ as to the dependability of the research. Kellaghan *et al.* (1996) commented that the results of experimental studies are not clear-cut and findings vary considerably with circumstances.

The meta-analysis by Deci *et al.* (1999) of 128 studies of the effects of extrinsic rewards on intrinsic motivation appear to show clearly that such rewards under-

mined intrinsic motivation across a wide range of activities, populations and types of reward. However, Hidi (2000) challenged these conclusions, pointing out that they were drawn from studies only relating to activities that were interesting, excluding uninteresting tasks. From their review of research on the role of interests and goals on achievement, Hidi and Harackiewicz (2000) concluded that the dichotomy between intrinsic 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.

The Review Methodology

Funding for this review was provided by the Nuffield Foundation and by the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre). The EPPI-Centre, established in 1999 with funding from the then Department for Education and Employment (DfEE—now Department for Education and Skills, DfES) in England, supports groups in undertaking systematic reviews of research in education to inform policy and practice. Its aim is to provide, in the education sector, a resource that gives policy makers and practitioners access to constantly up-dated results from synthesising research evidence. As a condition of funding, reviews are undertaken by groups, using systematic procedures, described later, which involve precise specification of the review parameters. The review reported here was one of the first to be conducted in the UK using EPPI-Centre procedures and software. A group (the Assessment and Learning Research Synthesis Group—ALRSG) was set up to steer reviews in assessment, whilst the review was carried out by the authors of this paper. Since these procedures represent a departure from those of narrative reviews, it is important to explain them at the start.

An aim of the EPPI-Centre is to create syntheses of relevant research that has been found at any one time that can be updated later. This is especially useful in a field of education such as assessment, where practice changes in response to frequent new policy initiatives and, more slowly, to feedback from research on the impact of policies. Specification and documentation of which studies have been reviewed and included in the synthesis is thus important, both for the interpretation of the findings at a particular date and for future work updating the review of the field.

The interest of the EPPI-Centre is to inform practice in school education. It funds reviews of research conducted with pupils of school age. Consequently, the search of the literature in this review was limited to those studies conducted with pupils aged 4 to 18. This had the effect of excluding studies of summative assessment in further and higher education, where the context and purpose of assessment is different from that in schools in certain significant respects (for example, the factors that give tests 'high stakes' in the school context).

The search for studies was completed in early 2001 and consequently the review did not include many studies published after 2000. This inevitably also excludes reference to important policy statements, such as the No Child Left Behind Act of 2001 in the USA. A further limitation was that the review included studies published

in English, found from searching data-bases and journals published in English. Although theoretically studies from all parts of the world could be included, it meant that studies published only in other languages were excluded. Moreover, although studies from several countries were read and included, our perspective as reviewers is inevitably influenced by our own background and current experience. The policy implications of the review findings, reported later, were drawn up in consultation with UK-based educators and policy makers, who identified what they saw as necessary change for UK policy and practice. Readers in other countries have to judge the feasibility and relevance of these implications for their own cultures. We are aware, for example, that the value of constructs such as intrinsic and extrinsic motivation in Chinese culture has been challenged (Watkins, 2000).

Within these parameters, the review included all types of studies. It did not give preference to randomised controlled trials; indeed in the contexts where testing is unavoidably part of students' experience, such study designs are often unrealistic. 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 on-going experience of students and not introduced by researchers in order to assess their impact. National 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 may mean that they have less weight in relation to implications for practice.

The review attempted to appraise the weight of evidence provided by the studies. Judgement of the overall weight that could be given to the evidence from a study was based on a combination of its 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. This is a review-specific judgement and does not represent a view of the quality of a study in its own right.

Procedures

The Review Questions

The first step in the systematic procedures employed in this review was to identify a review question at an appropriate level of specificity. The specification of the review question requires 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, 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 student 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. Thus the overall review question was expressed as:
What is the evidence of the impact of summative assessment and testing on students' motivation for learning? In order to achieve the aim 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?

Literature Search

The review question served as a framework in the search for studies. All the relevant electronic databases, journals held in accessible libraries and those on-line (which were very limited at the time of this review) were searched, citations in earlier reviews and in obtained papers were followed up and personal contacts used to obtain further references. This step, as all others of the review, was fully documented, recording, for example, dates of journals that were hand-searched and procedures for searching data-bases, so that the extent of the search was made explicit and the review can be updated later by reference to studies not included to date. The number of studies relevant to the review question found in this way was 183. Details of these, including abstracts, were entered into a data base. A list of these studies can be found in the full report of the review ⟨http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page = /reel/review_groups/assessment/review_one.htm⟩.

Applying Inclusion and Exclusion Criteria and Key-words

Before obtaining the full text of the studies, exclusion and inclusion criteria were applied to the abstracts. Studies were included if they were written in English, reported a study of a programme of summative testing or assessment involving students between the ages of 4 and 18, and reported on some aspect of motivation included in the meaning discussed earlier. The full texts of the 104 studies meeting these criteria were then obtained and read. Twenty-four studies were excluded at this stage due to mismatch between abstract and content or because they were not empirical studies. The next step was to describe the remaining empirical studies in

terms of a set of key-words, relating, for example, to their source, study type, age range and type of outcome reported. To check reliability in applying key-words, 30 studies were key-worded by two people. Agreement was considerable and differences helped in defining terms. Key-wording 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 it was re-coded as excluded. Sixty-one studies were not empirical studies but were reviews or were of sufficient relevance to be placed in a separate database labelled for use in background discussion and possible guidance in relation to recommendations.

Final Selection of Studies

At this point details of the included studies were discussed by the review group (ALRSG) and decisions made about a few studies that were borderline. Thus the final identification of a smaller number of studies (19) through this process ensures that attention is 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 is avoided.

Extraction and Evaluation of Evidence from the Studies

Data extraction was carried out using the *Guidelines for Extracting Data and Assessing Quality of Primary Studies in Educational Research, Version 0.94* (EPPI Reviewer—see website details above, p. 178). 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 on-line and off-line. Data were extracted from each study by at least two reviewers who then compared responses and reconciled differences. The process of extracting data from a study could take from four to six hours, depending on the length and complexity of the report.

Whilst all the 19 studies met the inclusion criteria and could be characterised using the general and specific key-words, they varied in design, methodology, instruments used and close relevance to the review questions. In order to ensure that conclusions were based on 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. The three aspects were: soundness of methodology of the study, as judged from the written report and revealed in the data extraction process; appropriateness of study type and design for answering the review questions; relevance of the topic focus of the study for answering the review questions. The judgements for these three aspects were combined into an overall weight to be given to the evidence in relation to the review focus.

Details of the final selection of 19 studies are set out in Table I, which gives for each one the evaluation of weight of evidence relevant to the review, the type of intervention, age group and country in which it was carried out and synthesis theme

TABLE I. Details of the 19 selected studies (see Appendix A for full references)

Study	Weight of evidence L = low; M = medium; H = high				Type of intervention	Age group	Country	'What I feel and think about myself as a learner'	'The energy I have for the task.'	'How I perceive my capacity to under-take the task.'
	Method ological quality	Relevance of study type	Relevance of study topic	Overall						
Benmansour (1999)	M	H	H	H	Naturalistic	High school	Morocco	×	×	×
Brookhart & Devoge (1999)	H	H	H	H	Naturalistic	High school	USA		×	×
Butler (1988)	H	H	H	H	Experimental	11 & 12 yrs	Israel		×	
Davies & Brember (1998)	H	H	H	H	Naturalistic	7 & 11 yrs	England	×		
Davies & Brember (1999)	M	H	H	H	Naturalistic	7 & 11 yrs	England	×		
Duckworth *et al.* (1986)	H	H	H	H	Naturalistic	High school	USA		×	×
Evans & Engelberg (1988)	H	H	H	H	Naturalistic	10–17 yrs	USA	×		
Ferguson & Francis (1979	H	H	M	M	Naturalistic	High school	England		×	
Gordon & Reese (1997)	M	H	M	M	Naturalistic	Elementary and High school	USA	×		
Hughes *et al.* (1986)	L	H	L	L	Experimental	11 yrs	USA		×	
Johnston & McClune (2000)	H	H	H	H	Naturalistic	11 & 12 yrs	Northern Ireland	×		×
Leonard & Davey (2001)	H	H	H	H	Naturalistic	11 yrs	Northern Ireland	×		
Little (1994)	L	H	M	M	Naturalistic	High school	England		×	
Paris *et al.* (1991)	L	H	M	M	Naturalistic	8–17 yrs	USA	×		
Perry (1998)	M	H	M	M	Naturalistic	6–9 yrs	Canada		×	
Pollard *et al.* (2000)	H	H	H	H	Naturalistic	5–11 yrs	England	×		
Reay & Wiliam (1999)	M	H	H	H	Naturalistic	11 yrs	England	×		
Roderick & Engel (2001)	M	H	M	M	Naturalistic	12 & 14 yrs	USA		×	
Schunk (1996)	H	H	H	H	Experimental	10 yrs	USA			×

TABLE II. Summary of features of the 19 selected studies

Design types		Type of intervention		Type of outcome (>1 per study)		Country of origin		Overall evidence weighting	
Outcome:									
RCT*	3	National Curriculum tests:	4	Effort	9	Canada	1	High	12
Case control	3	State tests:	3	Self-efficacy	4	Israel	1	Medium	6
Post-tests	3	11 +		Self-esteem	7	Morocco	1	Low	1
Other design	4	(Northern Ireland)	2	Interest	3	UK	8		
Process:	3	Classroom		Attitude to tests	5	USA	8		
Descriptive:	3	assessment:	5	Test anxiety	3				
		16 + exams	2	Learning -					
		Experimental (feedback)	1	Disposition	3				
		Experimental (other)	2	Self-regulation	2				
				Self as a learner	4				

* Randomised Controlled Trials

to which it contributed. Table II summarises information about the design types and types of outcome reported.

Synthesis of 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 testing on motivation for learning 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 single dependent variable which can be measured as an outcome, since, as discussed earlier, motivation for learning is a complex human attribute that is thought to be evidenced by a range of variables, each of which have affective, conative and cognitive dimensions. Nor are summative procedures the only factor affecting this complex overarching concept. A simplified view of the relationship is attempted in Figure 1.

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 in each. These outcomes fell into three distinct and overarching 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.'
> (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 task of synthesising the studies, to answer the main review question was tackled through focusing on the impact of tests on students' motivation for learning, examined through these three overarching themes which are deemed to be integral to it.

Consultation

The final phase of the methodology was to present the findings in progress to a peer group drawn together by the 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

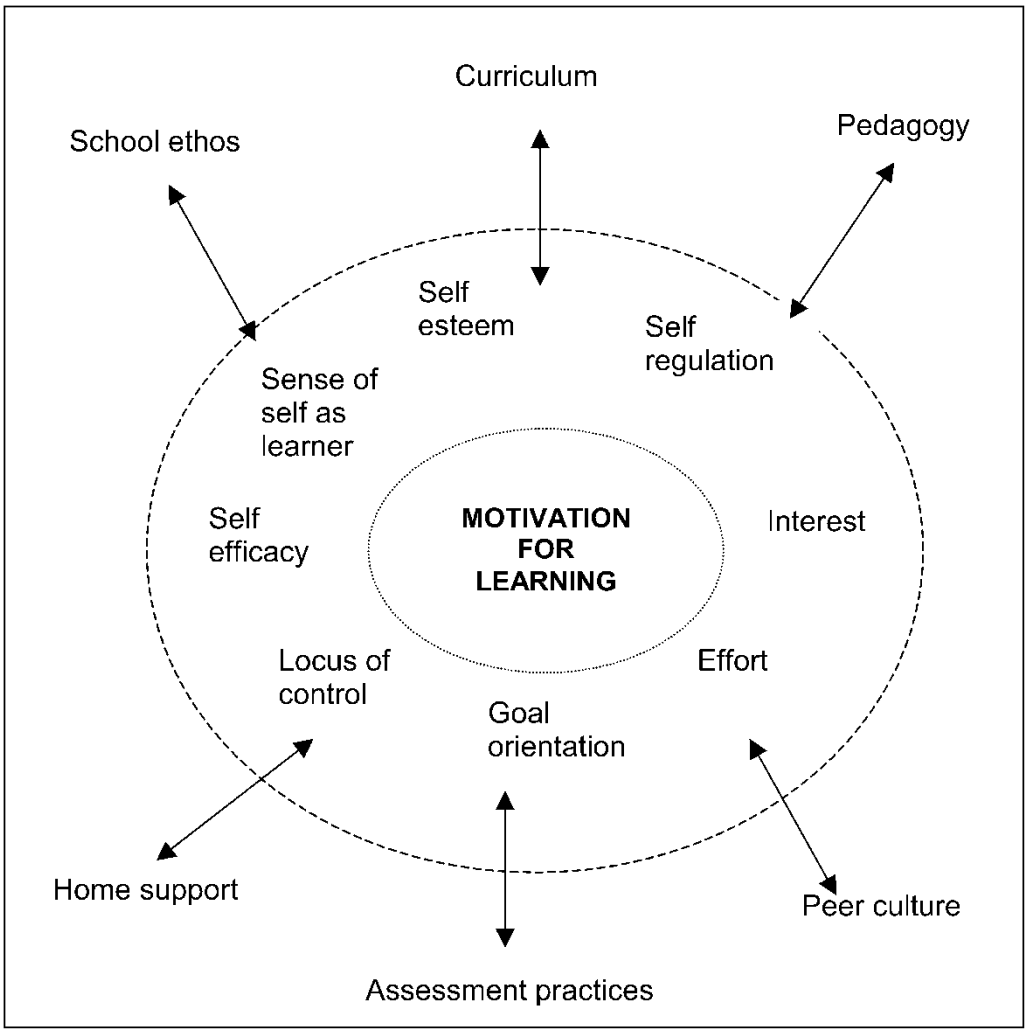

FIG. 1. Some of the variables relating to motivation and factors affecting them.

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 can be found on the ARG website ⟨www.assessment-reform-group.org.uk⟩.

Findings: evidence of impact on motivation for learning

The results of synthesising the review findings relating to the overall review question are given here in terms of the three themes identified above. The studies providing evidence for each of these are indicated in Table I.

'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.

Self-esteem

Two studies concerned the Northern Ireland end of primary school selection examination (known as the 11+ tests). 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+ tests.

Johnston and McLune (2000) 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 & Robinson, 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 and 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 gave priority to sequential processing and linked 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 (2001), funded by Save the Children, was specifically designed to reveal and publish students' views on the 11+ tests. 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 effort 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.

The impact of national tests in England and Wales was the subject of several studies. These tests were introduced in the 1988 Education Reform Act in England and Wales. A key part of this Act was the introduction of national tests for children in Years 2, 6 and 9 (ages 7, 11 and 14), phased in from 1989. The tests were designed to indicate achievement of individual students in terms of progressive levels (initially 1 to 10 and later modified to 1 to 8), the performance at each level being defined by achievement criteria. The levels are used to record and report individual progress but the tests results have also been used to set targets for and monitor the performance of schools, with consequent high stakes for the teachers.

From a small-scale study of a year 6 class in a London primary school in the term before the Year 6 (end of primary) national tests were taken, Reay and Wiliam (1999) reported perceptions of self-worth resulting from tests similar to those found by Leonard and Davey (2001). 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 made 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 & Wiliam, 1999, p. 346)

Two reports by Davies and Brember (1998, 1999) described results of an eight-year study of primary school children in England. Using the Lawseq question-

naire as a measure of self-esteem, they followed changes in the self-esteem of successive cohorts of Year 2 (age 7) and Year 6 (age 11) students over a period of eight years, starting two years before the National Tests were introduced at Year 2. 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 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 suggest that the initial drop in self-esteem was related to the circumstances surrounding the introduction of the tests for Year 2 children. Not only were these first tests complex, but 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. Once the national tests were simplified and teachers settled to a new regime, the Year 2 students' self-esteem rose. For the Year 6 students the tests did not begin until four years after the first Year 2 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 tests was Davies and Brember's (1998, 1999) finding that for pre-national test cohorts there was no correlation between self-esteem and achievement as measured by standardised tests in mathematics and reading. Post-national 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 national 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 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.

Studies by Gordon and Reese (1997) and Paris *et al.* (1991) both report on the impact of state mandated tests in the USA on the self-esteem of higher and lower achieving students. The differential impact of testing on low achieving students emerged in Gordon and Reese'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 and Test Anxiety

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 grades 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 (1988) 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. Pollard *et al.* (2000) followed a cohort of students, who were the first to be tested in Year 2, throughout their primary school. They collected data by questionnaire, interview, field notes and structured class observations and students' bubble cartoon completions. By the time the cohort reached Year 6, 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 (national 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' (p. 220).

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, the researchers suggested 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.* (2000) suggested 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. They concluded that assessment had a severely reduced role in helping learning and became concerned only with achievement as measured by testing, and there was evidence that students were all too aware of this.

Leonard and Davey (2001) reported that students' reactions to the Northern Ireland 11+ tests, with their explicit high stakes for the students' futures, were particularly strong. 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 Pollard *et al.* (2000) study, the initial excitement and novelty of taking practice tests soon wore off. Leonard and Davey (2001) found that students across all grade levels tended to be highly critical of the 11+ and wanted it to be abolished. Given that selection was inevitable, they favoured instead continuous assessment by the teacher

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 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 de-motivated, 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 test levels. Pollard *et al.* also concluded that students incorporated their teacher's evaluation of them into the construction of their identity as learners.

'The Energy I Put into the Task'

Nine studies were relevant to this outcome. Four of these provided high weight evidence, four provided medium weight evidence and one (not discussed) was judged to have only low weight in relation to the review questions (see Table I).

Feedback emerged from three studies as a significant factor influencing willingness to invest effort in a particular task. In one of these, 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 interviewed about their perceptions of their likely performance.

Students obtained feedback directly from their previous performance on similar tasks or from the teacher. Their judgements of their ability to succeed in particular assessments, such as spelling tests, was based on previous experience in spelling tests. 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 the 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.* (1986) 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:

1. Comments only: feedback consisted of one sentence, which related specifically to the performance of the individual child.
2. Grades only: these were based on the scores after conversion to follow a normal distribution with scores ranging from 40 to 99.
3. 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 score 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. The study of 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 concluded that promoting task involvement by giving task related, non-ego-involving, feedback may promote the interest and performance of most students.

Roderick and Engel (2001) reported the impact of a quite different approach to encouraging effort, by using the threat of consequences of failing tests. This study was the only one of the 19 that involved large proportions of minority students. 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 investigated the impact of this policy on 6th and 8th 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 (2001), drawing on questions from the base line interviews to code work effort, put students into four groups: those who were working harder in school as a result of the intervention (53% of the students); those working harder but outside of school, supported by other adults (9%); those who were 'worrying but not working' (34%); and those who were the most highly skilled in the sample and had already met targets in at least one subject (4%). Across the groups there were differences in age, gender and race. Eighth graders worked harder than 6th graders, males less than females and Latinos were 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 having 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 testing as a negative incentive means that some students will fail, and these will be the most vulnerable. However, an important finding is that schools can, by giving the kind of help described for the supporting school, raise students' achievement. The authors claimed that tests on their own, without this kind of support, do not raise achievement.

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). Students in three classes that were judged as being high in encouraging 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 (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' (p. 723).

Perry's (1998) findings 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' (p. 13).

'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.

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

earlier work 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 Pollard *et al.* (2000) also report, 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.

Working with high school students, Duckworth *et al.* (1986) 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

Johnston and McClune's (2000) study of the selection test for secondary schools in Northern Ireland, outlined on page 184, 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 processing' 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), in two linked experimental studies, explored self-regulatory processes among children who were learning mathematics. In both studies, two groups of students were randomly assigned to work under either a learning goal or a performance goal ethos. For the learning goal groups, the teacher introduced the

task, on manipulating fractions, by saying, 'While you are working it helps to keep in mind what you're trying to do', and went on: 'You'll be trying to learn how to solve fraction problems where the denominators are the same and you have to add the numerators'. For the performance goal groups the teacher gave the same first part of the instruction but did not go on to mention the explicit learning. For all the groups, the teacher asked the students to repeat the instructions to ensure they made sense to them. Thus the author claimed that, although there appeared to be a very small difference between the treatment of the groups, the particular instructions were registered by the students. 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, which were implemented in 45-minute instruction 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 than did those with performance goals and no self-evaluation. Self-evaluation scores for performance goals and for learning goals 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.

Benmansour's (1999) study, outlined on page 188, explored Moroccan students' perceived motivational orientations, self-efficacy, test anxiety and strategies used 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 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.

Findings: effect of characteristics of students and conditions of testing

Here we draw together information about the differential impact relating to age, level of achievement and gender of students and about the conditions that affect impact, from the studies as indicated in Table III.

TABLE III. 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 testing
Benmansour (1999)	H			×	×
Brookhart & Devoge (1999)	H				×
Butler (1988)	H		×		×
Duckworth *et al.* (1986)	H				×
Evans & Engelberg (1988)	H	×	×	×	
Ferguson and Francis (1979)	M			×	
Gordon & Reese (1997)	M		×		×
Johnston & McClune (2000)	H			×	
Leonard & Davey (2001)	H		×		×
Little (1994)	M				×
Paris *et al.* (1991)	M	×	×		
Perry (1998)	M				×
Pollard *et al.* (2000)	H	×	×		×
Reay & Wiliam (1999)	H		×	×	×
Roderick & Engel (2001)	M	×	×		×

Key: H = high weight of evidence M = medium weight of evidence L = low weight of evidence

Age of Students

Two studies indicated that reactions to grades, attribution and goal orientation vary with students' age. Evans and Engelberg's (1988) study of teachers' classroom marking or grading, showed that older students (that is, age 11 and above) were 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 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.

The findings of Paris *et al.* (1991) suggest that lower achieving older students were more likely to minimise effort and respond to test items randomly or by guessing than younger ones. Thus tests have progressively less validity for these children. However, under threat of serious consequences for not reaching a required level, eighth graders were more likely to work harder than sixth graders (Roderick & 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 of summative classroom assessment show that high achieving students are generally less affected by grading than low achievers (Paris *et al.*, 1991; Pollard *et al.*, 2000). 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 & 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 & Davey, 2001). All students were aware of repeated practice tests and the narrowing of the curriculum and only those confident of success enjoy the tests (Reay & 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 actual understanding and skills. 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 (Paris *et al.*, 1991; Pollard *et al.*, 2000).

Evidence on the differential impact of testing on low achieving students emerged in two studies of state-mandated tests in the USA. Gordon and Reese's (1997) exploration of the reactions of teachers in the State of Texas to the TAAS found a

strong perception that tests lowered the self-esteem of students 'at risk'. Similarly, Paris *et al.* (1991) found from information collected about the Michigan State mandated tests, that high achievers had more positive self-perceptions than low achievers.

Several studies show evidence 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. But there is evidence 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 & Engel, 2001).

Gender

Differences in learning dispositions of boys and girls were found to have particular importance in classrooms that favour certain approaches to learning. Johnston and McClune (2000) 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. 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 (Benmansour, 1999; Evans & Engelberg, 1988; Reay & 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 & Engelberg, 1988).

Ferguson and Francis (1979) studied modes of examination and motivation of students taking the GCE 'O' level examination in English. At the time of their study candidates could be entered either for an examination or for continuous course assessment by teachers. Although there were some differences in attitude towards the subject resulting from mode of examination, these were not significant. The significant differences in attitude resulted from gender and to a lesser extent place of study (school or college).

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 (Gordon & Reese, 1997; Perry, 1998; Pollard *et al.*, 2000; Reay & Wiliam, 1999; Roderick & Engel, 2001).

The importance of self-efficacy in supporting student effort and achievement is a thread in several studies. Feedback has a central role in this since self-efficacy is judged from performance in previous tasks of the same kind (Brookhart & 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 (Brookhart & DeVoge, 1999; Butler, 1988). Goal-orientation, effort and interest are all interconnected. Benmansour (1999) reported that students who are performance orientated have less interest in the task *per se* and that 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.

Perry (1998) found that 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 is correct or not. In other classrooms students evaluated their work by reference to surface features, such as whether it was neat, well presented and 'right', as was also found by 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 & Davey, 2001; Little, 1994).

There is a strong basis of evidence that community pressure is brought to bear on schools for high scores (Gordon & Reese, 1997; Reay & 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 & Davey, 2001). For many students this increases students' anxiety even though they recognised their parents as being supportive (Leonard & Davey, 2001; Reay & Wiliam, 1999).

Findings: impact on teachers and teaching

The following findings were brought together from those studies that, in addition to reporting impact on students' motivation, provided evidence of impact of testing on teaching and teachers. All seven of these studies pointed to very similar effects of high stakes summative assessment.

Johnston and McClune (2000) found that the existence of external tests has a constricting effect on the curriculum and on teaching methods. Reay and Wiliam (1999) reported that emphasis in teaching was based 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 was given to subjects not tested. Areas particularly neglected are those related to creativity and personal and social development (Gordon & Reese, 1997; Leonard & Davey, 2001).

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 & Reese, 1997; Leonard & Davey, 2001). 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' (p. 364). 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. Not only is the scope and depth of learning seriously undermined, but this also affects the validity of the tests, for they no longer indicate 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 & 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. Pollard *et al.*'s (2000) study, covering the introduction of the national tests 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 dealing with students as they get older. 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 & McClune, 2000) and also show how readily students pick up from their teacher the signs of what is valued and will gain approval. Thus, as teachers become more performance-centred, students pick up the criteria being used and judge their own work accordingly (Pollard *et al.*, 2000). There is evidence that teachers can influence children's self-assessment to focus on learning processes (e.g. 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.

Roderick and Engel (2001) concluded 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.

Findings: reducing the negative and increasing the positive impact

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 & Engel, 2001). Reay and Wiliam (1999) also note that threats to schools posed by poor national test results put teachers under pressure to increase scores by whatever means, regardless of the longer term impact on students' learning.

This and other evidence points to the following effect of raising the stakes:

- Increase in test anxiety (Benmansour, 1999; Leonard & 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 & Brember, 1998; Leonard & Davey, 2001).
- Teaching being focused on the content of the tests and teaching methods confined to transmission modes which favour sequential learning styles (Johnston & 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 & Wiliam, 1999).

These effects are similar in high and low achieving schools (Johnston & McClune, 2000; Pollard *et al.*, 2000) and apply equally to high and low achieving students (Gordon & Reese, 1997).

Reducing the Negative Impact

All but two of the selected studies provided some information relating to possible causes of tests affecting motivation and by implication provide suggestions for reducing the negative and increasing the positive impact of tests. These are summarised briefly here and taken up later in discussion of implications for assessment policy and practice.

A number of findings point to practices that, if reduced or curtailed, would decrease the negative impact of tests. These include focusing teaching on the test content, training students to pass the tests and using class time for repeated practice tests (Gordon & Reese, 1997; Johnston & McClune, 2000; Leonard & Davey, 2001; Paris *et al.*, 1991; Reay & Wiliam, 1999).

More positive action is also suggested. This includes

- Promoting learning goal orientation rather than performance goal orientation (Brookhart & DeVoge, 1999; Roderick & Engel, 2001; Schunk, 1996).
- Cultivating intrinsic interest in the subject and put less emphasis on grades (Benmansour, 1999) but make grading criteria explicit (Evans & Engelberg, 1988).
- Emphasising teaching approaches that encourage collaboration among students and cater for a range of teaching styles (Johnston & McClune, 2000; Pollard *et al.*, 2000; Reay & Wiliam, 1999).
- Explaining the reasons for, and the implications of, tests (Leonard & Davey, 2001; Pollard *et al.*, 2000).
- Providing feedback to students about their performance in a form that is non-ego-involving and non-judgemental (Brookhart & DeVoge, 1999; Butler, 1988) and helping students to interpret it (Duckworth *et al.*, 1986).
- Broadening the range of information used in assessing the attainment of individual students (Reay & Wiliam, 1999) and broadening the base of information used in evaluating the effectiveness of schools (Gordon & Reese, 1997).

Increasing the Positive Impact

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 the studies indicate action that would enable summative testing and assessment to take a positive role in students' learning:

- Ensuring that the demands of the tests are consistent with the expectations of teachers and the capabilities of the students (Duckworth *et al.* 1986).
- Involving students in decisions about testing (Leonard & Davey, 2001; Little, 1994).
- Developing students' self-assessment skills and use of learning rather than performance criteria (Pollard *et al.*, 2000; Schunk, 1996).

- Developing a constructive and supportive school ethos in relation to tests (Duckworth *et al.*, 1986).
- Using assessment to convey a sense of learning progress to students (Duckworth *et al.*, 1986; Roderick & Engel, 2001).
- Supporting low-achieving students' self-efficacy by making learning goals explicit and showing them how to direct effort in learning (Roderick & Engel, 2001).
- Creating a classroom environment that promotes self-regulated learning (Perry, 1998).

Implications for Assessment Policy and Practice Identified through Discussion of the Findings

This review was funded and conducted for the explicit purpose of identifying dependable findings of relevance to assessment policy and practice. In drawing out implications, 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. The conference was a planned part of the procedures of the review (see p. 182) and the outcome of the deliberations are included in the implications discussed here. The conference considered the findings from the review in the context not just of summative assessment but against the wider background of assessment in education, particularly in the UK.

Implications for Practice

Many of the findings summarised have clear messages for how the negative impact of tests on motivation for learning can be minimised. In some cases these refer to practices that should be ended as far as possible. In particular they suggest avoiding drill and practice for tests, de-emphasising tests by using a range of forms of classroom assessment and recognising the limitations of tests, preventing the content and methods of teaching from being limited by the form and content of tests and taking steps to prevent children being faced with tests in which they are unlikely to succeed. These may seem unrealistic to some who feel unable to resist the grip of current testing regimes, but they should still be recognised as goals to pursue as conditions allow.

However, rather than indicate only what should be avoided, there are more positive messages for action that teachers and schools can take to ensure that the benefits of summative assessment can be had without negative impact on students' motivation for learning. The following were identified:

a. Promote and engage in professional development that emphasises learning goals and learner-centred teaching approaches to counteract the narrowing of the curriculum.

b. Share and emphasise with students learning goals, not performance goals, and provide feedback to students in relation to these goals.
c. Develop and implement 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, for example, by McCombs (1999).
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.

Implications for Assessment Policy

Teachers work within the structures and limitations set by schools, by district or local education requirements and by national policies. There are limits to the action they can take to use assessment effectively to help their students' learning, and yet they are the only ones whose actions directly affect students. Governments are recognising in their education policies the importance of promoting continued learning throughout life, as needed by citizens of a world in which the pace of change is not just continuing but is accelerating. Evidence from this review, however, suggests that current testing practices are detrimental to, rather than encouraging of, the attitudes and energy for learning needed for lifelong learning.

Some of the directions in which change is needed emerged from the discussion of the review findings at the consultation conference. The participants drew on their experiences and knowledge of other research and practice, thus several of these points go beyond the evidence base of the research review. They are developed further in an Assessment Reform Group pamphlet (ARG, 2002).

A key point to policy makers is to recognise that current high stakes testing is failing to provide valid information about students' attainment for a number of reasons. For example, the tests are too narrowly focused to provide information about students' attainment and the consequences of teaching to the tests mean 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 these skills and understanding.

There should be more emphasis placed on 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, in the findings of the OECD/PISA study (OECD, 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. The recognition of these valued outcomes could be conveyed, for example, by requiring that criteria used in school evaluation, including self-evaluation, make explicit reference to a full range of subjects and include spiritual, moral, social and cultural as well as cognitive aims and an appropriate variety of teaching methods and learning outcomes. The current human and financial resources devoted to test development could be used to create assessment systems that enable all valued outcomes of education, including creativity and learning to learn to be assessed.

It was noted that alternatives to tests to give summative information about individual students, avoiding the negative impact on students, could be found in programmes of testing students when their teachers judge them to be ready to show their achievement at a certain level. For tracking national standards, more valid and useful information, from a wider range of test forms and items, can be gained by sampling students rather than testing whole cohorts.

It was emphasised that assessment policy makers should be aware of the real cost of current practice, including teaching time taken up for testing and practice testing and adding to teachers' workloads, in addition to the cost of the tests and their development.

Finally the policy of setting targets based only on test results was identified as a key factor in raising the stakes to the point where test testing begins to act in opposition to the intentions of reform. Interestingly the chief inspector for schools in England has reported 'a very real concern that the innovation and reform that we need to see in our schools may be inhibited by an over-concentration on targets' (Bell, 2003).

Conclusion

One of the main outcomes of the research review is to draw attention to the small number of studies that were found to offer dependable evidence to address the question posed in this review. 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 outcomes of assessment is very small by comparison. We have argued that there are important reasons for serious attention to motivation for learning as an outcome of education. We have also discussed the complexity of the concept of motivation for learning and indicated that it can be discouraged unwittingly by assessment and testing practices. It is not the role of this paper to suggest how to promote motivation, but the review has hopefully pointed out some of the actions and conditions that impact both positively and negatively on it.

REFERENCES (not including studies listed in Appendix A)

AMERICAN PSYCHOLOGICAL ASSOCIATION (1997) *Learner-Centred Principles: a framework for school reform and redesign* (Washington DC, American Psychological Association).

AMES, C. (1990a) Motivation: what teachers need to know, *Teachers College Record*, 91, pp. 409–421.

AMES, C. (1990b) Developing a learning orientation. Paper presented at annual meeting of the AERA, Boston, 16–20 April.

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

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

BELL, D. (2003) Reporting England—Speech to the City of York Council's annual education conference. February. OfSTED News ⟨www.ofsted.gov.uk/news⟩.

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

BLACK, P. & WILIAM, D. (1998) Assessment and classroom learning, *Assessment in Education*, 5 (1), pp. 7–74.

BROADFOOT, P. & POLLARD, A. (2000) The changing discourse of assessment policy: the case of English primary education, in: A. FILER (Ed.) *Assessment: social practice and social product* (London, Falmer Press).

CLARKE, M., MADAUS, G. F., HORN, C. J. & RAMOS, M. A. (2000) Retrospective on educational testing and assessment in the 20th century, *Journal of Curriculum Studies*, 32 (2), pp. 159–181.

CROOKS, T. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58, pp. 438–481.

DECI, E. L. & RYAN R. M. (1985) *Intrinsic Motivation and Self-determination in Human Behavior* (Plenum, New York).

DECI, E. L., KOESTNER, R. & RYAN, R. M. (1999) A meta-analysis review of experiments examining the effects of extrinsic rewards on intrinsic motivation, *Psychological Bulletin*, 125, pp. 627–688.

DWECK, C. S. (1992) The study of goals in psychology, *Psychological Science*, 3, pp. 165–167.

GROLNICK, W. S. & RYAN, R. M. (1987) Autonomy in children's learning: an experimental and individual difference investigation, *Journal of Personality and Social Psychology*, 52, pp. 890–898.

HIDI, S. (2000) An interest researcher's perspective: the effects of extrinsic and intrinsic factors on motivation, in: C. SANSOME & J. M. HARACKIEWICZ (Eds) *Intrinsic and Extrinsic Motivation: the search for optimal motivation and performance* (New York, Academic Press).

HIDI, S. & HARACKIEWICZ, J. M. (2000) Motivating the academically unmotivated: a critical issue for the 21st century, *Review of Educational Research*, 70 (2), pp. 151–179.

JOHNSTON, C. (1996) *Unlocking the Will to Learn* (Thousand Oaks, CA, Corwin Press).

KELLAGHAN T., MADAUS G. & RACZEK, A. (1996) *The Use of External Examinations to Improve Student Motivation* (Washington DC, AERA).

KOHN, A. (1993) *Punished by Rewards* (Boston, MA, Houghton Mifflin).

KOHN, A. (2000) *The Case Against Standardized Testing* (Portsmouth, NH, Heinemann).

KORETZ, D. (1988) Arriving at Lake Wobegon: are standardised tests exaggerating achievement and distorting instruction? *American Educator*, 12 (2), pp. 8–15.

KORETZ, D., LINN, R. L, DUNBAR, S. B. & SHEPARD, L. A. (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, IL, 3–7 April.

LINN, R. (2000) Assessments and accountability, *Educational Researcher*, 29, pp. 4–16.

MADAUS, G. & CLARKE, M. (1999) 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, 4 December, 1998. Paper revised May 1999.

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

MCCOMBS, B. L. (1999) *Learner-Centred Classroom Practices.* Available from the author, University of Denver Research Institute, Denver, Colorado.

MCCOMBS, B. L. & WHISLER, J. (1997) *The Learner Centred Classroom and School* (San Francisco, CA, Jossey-Bass).

MCDONALD, A. (2001) The prevalence and effects of test anxiety in school children, *Educational Psychology*, 21, pp. 89–101.

MCNEIL, L. & VALENZUELA, A. (1998) The harmful effects of the TAAS system of testing in Texas: beneath the accountability rhetoric, High Stakes K-12 Testing Conference, Harvard University, 4 December, 1998.

NORWICKI, S. & STRICKLAND, B. (1973) A locus of control scale for children, *Journal of Consulting and Clinical Psychology*, 40, pp. 148–155.

OECD (2001) *Knowledge and Skills for Life. First results from PISA 2000* (Paris, OECD).

OSBORN, M., MCNESS, E., BROADFOOT, P., POLLARD, A. & TRIGGS, P. (2000) *What Teachers Do: changing policy and practice in primary education* (London, Continuum).

PROFESSIONAL ASSOCIATION OF TEACHERS (2000) Press release 06/01/00. See www.pat.org.uk

RESNICK, L. B. & NOLAN, K. L. (1995) Standards for education, in: D. RAVITCH (Ed.) *Debating the Future of American Education: do we need national standards and assessment?* (Washington DC, Brookings Institution).

SCHOEN, H. L., FEY, J. T., HIRSCH, C. R. & COXFORD, A. E. (1999) Issues and options in math wars, *Phi Delta Kappan*, February, pp. 444–453.

SCHUNK, D. (1991) Self-efficacy and academic motivation, *Educational Psychologist*, 26, pp. 207–231.

STIGGINS, R. (2001) *Student-Involved Classroom Assessment* (3rd edn) (Upper Saddle River, NJ, Merrill Prentice Hall).

WATKINS, D. (2000) Learning and teaching: a cross-cultural perspective, *School Leadership and Management*, 20 (2), pp. 161–173.

Appendix A: List of the 19 studies

1. BENMANSOUR, N. (1999) Motivational orientations, self-efficacy, anxiety and strategy use in learning high school mathematics in Morocco, *Mediterranean Journal of Educational Studies*, 4, pp. 1–15.
2. BROOKHART, S. & DEVOGE, J. (1999) Testing a theory about the role of classroom assessment in pupil motivation and achievement, *Applied Measurement in Education*, 12, pp. 409–425.
3. 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, pp. 1–14.
4. 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, pp. 365–375.
5. 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, pp. 145–157.
6. DUCKWORTH, K., FIELDING, G. & SHAUGHNESSY, J. (1986) *The Relationship of High School Teachers' Class Testing Practices to Pupils' Feelings of Efficacy and Efforts to Study* (Portland, OR, Oregon University).
7. EVANS, E. & ENGELBERG, R. (1988) Pupils' perceptions of school grading, *Journal of Research and Development in Education*, 21, pp. 44–54.

8. 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), pp. 231–239.
9. GORDON, S. & REESE, M. (1997) High stakes testing: worth the price? *Journal of School Leadership*, 7, pp. 345–368.
10. 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, pp. 660–667.
11. 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, in: *The Effects of the Selective System of Secondary Education in Northern Ireland*: Research Papers Volume II (Bangor, Co. Down, Department of Education).
12. LEONARD, M. & DAVEY, C. (2001) *Thoughts on the 11 Plus* (Belfast, Save the Children Fund).
13. LITTLE, A. (1994) Types of assessment and interest in learning: variation in the south of England in the 1980s, *Assessment in Education*, 1, pp. 201–222.
14. PARIS, S., LAWTON, T., TURNER, J. & ROTH, J. (1991) A developmental perspective on standardised achievement testing, *Educational Researcher*, 20, pp. 12–20.
15. PERRY, N. (1998) Young children's self-regulated learning and contexts that support it, *Journal of Educational Psychology*, 90, pp. 715–729.
16. POLLARD, A., TRIGGS, P., BROADFOOT, P., MCNESS, E. & OSBORN, M. (2000) *What Pupils Say: changing policy and practice in primary education* (London, Continuum).
17. REAY, D. & WILIAM, D. (1999) 'I'll be a nothing': structure, agency and the construction of identity through assessment, *British Educational Research Journal*, 25, pp. 343–354.
18. 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, pp. 197–228.
19. SCHUNK D. (1996) Goal and self-evaluative influences during children's cognitive skill learning, *American Educational Research Journal*, 33, pp. 359–382.