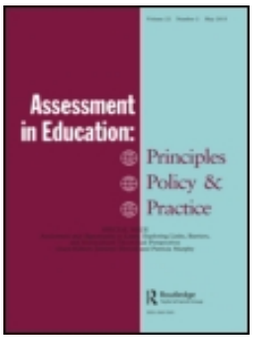

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: www.tandfonline.com/journals/caie20

Assessment and Classroom Learning

Paul Black & Dylan Wiliam

To cite this article: Paul Black & Dylan Wiliam (1998) Assessment and Classroom Learning, Assessment in Education: Principles, Policy & Practice, 5:1, 7-74, DOI: 10.1080/0969595980050102

To link to this article: https://doi.org/10.1080/0969595980050102

Assessment and Classroom Learning

PAUL BLACK & DYLAN WILIAM
School of Education, King's College London, Cornwall House, Waterloo Road, London SE1 8WA, UK

ABSTRACT *This article is a review of the literature on classroom formative assessment. Several studies show firm evidence that innovations designed to strengthen the frequent feedback that students receive about their learning yield substantial learning gains. The perceptions of students and their role in self-assessment are considered alongside analysis of the strategies used by teachers and the formative strategies incorporated in such systemic approaches as mastery learning. There follows a more detailed and theoretical analysis of the nature of feedback, which provides a basis for a discussion of the development of theoretical models for formative assessment and of the prospects for the improvement of practice.*

Introduction

One of the outstanding features of studies of assessment in recent years has been the shift in the focus of attention, towards greater interest in the interactions between assessment and classroom learning and away from concentration on the properties of restricted forms of test which are only weakly linked to the learning experiences of students. This shift has been coupled with many expressions of hope that improvement in classroom assessment will make a strong contribution to the improvement of learning. So one main purpose of this review is to survey the evidence which might show whether or not such hope is justified. A second purpose is to see whether the theoretical and practical issues associated with assessment for learning can be illuminated by a synthesis of the insights arising amongst the diverse studies that have been reported.

The purpose of this Introduction is to clarify some of the key terminology that we use, to discuss some earlier reviews which define the baseline from which our study set out, to discuss some aspects of the methods used in our work, and finally to introduce the structure and rationale for the subsequent sections.

Our primary focus is the evidence about formative assessment by teachers in their school or college classrooms. As will be explained below, the boundary for the research reports and reviews that have been included has been loosely rather than tightly drawn. The principal reason for this is that the term formative assessment does not have a tightly defined and widely accepted meaning. In this review, it is to be interpreted as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the

0969-594X/98/010007-68

teaching and learning activities in which they are engaged.

Two substantial review articles, one by Natriello (1987) and the other by Crooks (1988) in this same field serve as baselines for this review. Therefore, with a few exceptions, all of the articles covered here were published during or after 1988. The literature search was conducted by several means. One was through a citation search on the articles by Natriello and Crooks, followed by a similar search on later and relevant reviews of component issues published by one of us (Black, 1993b), and by Bangert-Drowns and the Kuliks (Kulik *et al.*, 1990; Bangert-Drowns *et al.*, 1991a,b). A second approach was to search by key-words in the ERIC data-base; this was an inefficient approach because of a lack of terms used in a uniform way which define our field of interest. The third approach was the 'snowball' approach of following up the reference lists of articles found. Finally, for 76 of the most likely journals, the contents of all issues were scanned, from 1988 to the present in some cases, from 1992 for others because the work had already been done for the 1993 review by Black (see Appendix for a list of the journals scanned).

Natriello's review covered a broader field than our own. The paper spanned a full range of assessment purposes, which he categorised as certification, selection, direction and motivation. Only the last two of these are covered here. Crooks used the term 'classroom evaluation' with the same meaning as we propose for 'formative assessment'. These two articles gave reference lists containing 91 and 241 items respectively, but only 9 items appear in both lists. This illustrates the twin and related difficulties of defining the field and of searching the literature.

The problems of composing a framework for a review are also illustrated by the differences between the Natriello and the Crooks articles. Natriello reviews the issues within a framework provided by a model of the assessment cycle, which starts from purposes, then moves to the setting of tasks, criteria and standards, then through to appraising performance and providing feedback and outcomes. He then discusses research on the impact of these evaluation processes on students. Perhaps his most significant point, however, is that in his view, the vast majority of the research into the effects of evaluation processes is irrelevant because key distinctions are conflated (for example by not controlling for the quality as well as the quantity of feedback). He concludes by suggesting how the weaknesses in the existing research-base might be addressed in future research.

Crooks' paper has a narrower focus—the impact of evaluation practices on students—and divides the field into three main areas—the impact of normal classroom testing practices, the impact of a range of other instructional practices which bear on evaluation, and finally the motivational aspects which relate to classroom evaluation. He concludes that the summative function of evaluation—grading—has been too dominant and that more emphasis should be given to the potential of classroom assessments to assist learning. Feedback to students should focus on the task, should be given regularly and while still relevant, and should be specific to the task. However, in Crooks' view the 'most vital of all the messages emerging from this review' (p. 470) is that the assessments must emphasise the skills, knowledge and attitudes perceived to be most important, however difficult the technical problems that this may cause.

Like Natriello's review, the research cited by Crooks covers a range of styles and contexts, from curriculum-related studies involving work in normal classrooms by the students' own teachers, to experiments in laboratory settings by researchers. The relevance of work that is not carried out in normal classrooms by teachers can be called in question (Lundeberg & Fox, 1991), but if all such work were excluded, not only would the field be rather sparsely populated, but one would also be overlooking many important clues and pointers towards the difficult goal of reaching an adequately complex and complete understanding of formative assessment. Thus this review, like that of Natriello and more particularly that of Crooks, is eclectic. In consequence, decisions about what to include have been somewhat arbitrary, so that we now have some sympathetic understanding of the lack of overlap between the literature sources used in the two earlier reviews.

The processes described above produced a total of 681 publications which appeared relevant, at first sight, to the review. The bibliographic details for those identified by electronic means were imported (in most cases, including abstracts) into a bibliographic database, and the others were entered manually. An initial review, in some cases based on the abstract alone, and in some cases involving reading the full publication, identified an initial total of about 250 of these publications as being sufficiently important to require reading in full. Each of these publications was then coded with labels relating to its focus—a total of 47 different labels being used, with an average of 2.4 labels per reference. For each of the labelled publications, existing abstracts were reviewed and, in some cases modified to highlight aspects of the publication relevant to the present review, and abstracts written where none existed in the database. Based on a preliminary reading of the relevant papers, a structure of seven main sections was adopted.

The writing for each section was undertaken by first allocating each label to a section. All but one of the labels was allocated to a unique section (one was allocated to two sections). Abstracts of publications relevant to each section were then printed out together and each section was allocated to one of the authors so that initial drafts could be prepared, which were then revised jointly. The seven sections which emerged from this process may be briefly described as follows.

The approach in the section on Examples in evidence is pragmatic, in that an account is given first of a variety of selected pieces of research about the effectiveness of formative assessment, and then these are discussed in order to identify a set of considerations to be borne in mind in the succeeding—more analytic—sections. The next section on Assessment by teachers adds to the empirical background by presenting a brief account of evidence about the current state of formative assessment practice amongst teachers.

There follows a more structured account of the field. The next two sections deal respectively with the student perspective and the teachers' role. Whilst the section on Strategies and tactics for teachers focuses on tactics and strategies in general, the next section on Systems follows by discussing some specific and comprehensive systems for teaching in which formative assessment plays an important part. The section on Feedback is more reflective and theoretical, presenting an account, grounded in evidence, of the nature of feedback, a concept which is central to

formative assessment. This prepares the ground for a final section, on Prospects for the theory and practice of formative assessment, in which we attempt a synthesis of some of the main issues in the context of an attempt to review the theoretical basis, the research prospects and needs, and the implications for practice and for policy of formative assessment studies.

Examples in Evidence

Classroom Experience

In this section we present brief accounts of pieces of research which, between and across them, illustrate some of the main issues involved in research which aims to secure evidence about the effects of formative assessment.

The *first* is a project in which 25 Portuguese teachers of mathematics were trained in self-assessment methods on a 20-week part-time course, methods which they put into practice as the course progressed with 246 students of ages 8 and 9 and with 108 older students with ages between 10 and 14 (Fontana & Fernandes, 1994). The students of a further 20 Portuguese teachers who were taking another course in education at the time served as a control group. Both experimental and control groups were given pre- and post- tests of mathematics achievement, and both spent the same times in class on mathematics. Both groups showed significant gains over the period, but the experimental group's mean gain was about twice that of the control group's for the 8 and 9-year-old students—a clearly significant difference. Similar effects were obtained for the older students, but with a less clear outcome statistically because the pre-test, being too easy, could not identify any possible initial difference between the two groups. The focus of the assessment work was on regular—mainly daily—self-assessment by the pupils. This involved teaching them to understand both the learning objectives and the assessment criteria, giving them opportunity to choose learning tasks and using tasks which gave them scope to assess their own learning outcomes.

This research has ecological validity, and gives rigorously constructed evidence of learning gains. The authors point out that more work is required to look for long-term outcomes and to explore the relative effectiveness amongst the several techniques employed in concert. However, the work also illustrates that an initiative can involve far more than simply adding some assessment exercises to existing teaching—in this case the two outstanding elements are the focus on self-assessment and the implementation of this assessment in the context of a constructivist classroom. On the one hand it could be said that one or other of these features, or the combination of the two, is responsible for the gains, on the other it could be argued that it is not possible to introduce formative assessment without some radical change in classroom pedagogy because, of its nature, it is an essential component of the pedagogic process.

The *second* example is reported by Whiting *et al.* (1995), the first author being the teacher and the co-authors university and school district staff. The account is a review of the teacher's experience and records, with about 7000 students over a

period equivalent to 18 years, of using mastery learning with his classes. This involved regular testing and feedback to students, with a requirement that they either achieve a high test score—at least 90%—before they were allowed to proceed to the next task, or, if the score were lower, they study the topic further until they could satisfy the mastery criterion. Whiting's final test scores and the grade point averages of his students were consistently high, and higher than those of students in the same course not taught by him. The students' learning styles were changed as a result of the method of teaching, so that the time taken for successive units was decreased and the numbers having to retake tests decreased. In addition, tests of their attitudes towards school and towards learning showed positive changes.

Like the previous study, this work has ecological validity—it is a report of work in real classrooms about what has become the normal method used by a teacher over many years. The gains reported are substantial; although the comparisons with the control are not documented in detail, it is reported that the teacher has had difficulty explaining his high success rate to colleagues. It is conceded that the success could be due to the personal excellence of the teacher, although he believes that the approach has made him a better teacher. In particular he has come to believe that all pupils can succeed, a belief which he regards as an important part of the approach. The result shows two characteristic and related features—the first being that the teaching change involves a completely new learning regime for the students, not just the addition of a few tests, the second being that precisely because of this, it is not easy to say to what extent the effectiveness depends specifically upon the quality and communication of the assessment feedback. It differs from the first example in arising from a particular movement aimed at a radical change in learning provision, and in that it is based on different assumptions about the nature of learning.

The *third* example also had its origin in the idea of mastery learning, but departed from the orthodoxy in that the authors started from the belief that it was the frequent testing that was the main cause of the learning achievements reported for this approach. The project was an experiment in mathematics teaching (Martinez & Martinez, 1992), in which 120 American college students in an introductory algebra course were placed in one of four groups in a 2×2 experimental design for an 18-week course covering seven chapters of a text. Two groups were given one test per chapter, the other two were given three tests per chapter. Two groups were taught by a very experienced and highly rated teacher, the other two by a relatively inexperienced teacher with average ratings. The results of a post-test showed a significant advantage for those tested more frequently, but the gain was far smaller for the experienced teacher than for the newcomer. Comparison of the final scores with the larger group of students in the same course but not in the experiment showed that the experienced teacher was indeed exceptional, so that the authors could conclude that the more frequent testing was indeed effective, but that much of the gain could be secured by an exceptional teacher with less frequent testing.

By comparison with the first study above, this one has similar statistical measures and analyses, but the nature of the two regimes being compared is quite different. Indeed, one could question whether the frequent testing really constitutes formative

assessment—a discussion of that question would have to focus on the quality of the teacher–student interaction and on whether test results constituted feedback in the sense of leading to corrective action taken to close any gaps in performance (Ramaprasad, 1983). It is possible that the superiority of the experienced teacher may have been in his/her skill in this aspect, thus making the testing more effectively formative at either frequency.

Example number *four* was undertaken with 5-year-old children being taught in kindergarten (Bergan *et al.*, 1991). The underlying motivation was a belief that close attention to the early acquisition of basic skills is essential. It involved 838 children drawn mainly from disadvantaged home backgrounds in six different regions in the USA. The teachers of the experimental group were trained to implement a measurement and planning system which required an initial assessment input to inform teaching at the individual pupil level, consultation on progress after two weeks, new assessments to give a further diagnostic review and new decisions about students' needs after four weeks, with the whole course lasting eight weeks. The teachers used mainly observations of skills to assess progress, and worked with open-style activities which enabled them to differentiate the tasks within each activity in order to match to the needs of the individual child. There was emphasis in their training on a criterion-referenced model of the development of understanding drawn up on the basis of results of earlier work, and the diagnostic assessments were designed to help locate each child at a point on this scale. Outcome tests were compared with initial tests of the same skills. Analysis of the data using structural equation modelling showed that the pre-test measures were a strong determinant of all outcomes, but the experimental group achieved significantly higher scores in tests in reading, mathematics and science than a control group. The criterion tests used, which were traditional multiple-choice, were not adapted to match the open child-centred style of the experimental group's work. Furthermore, of the control group, on average 1 child in 3.7 was referred as having particular learning needs and 1 in 5 was placed in special education; the corresponding figures for the experimental group were 1 in 17 and 1 in 71.

The researchers concluded that the capacity of children is under-developed in conventional teaching so that many are 'put down' unnecessarily and so have their futures prejudiced. One feature of the experiment's success was that teachers had enhanced confidence in their powers to make referral decisions wisely. This example illustrates again the embedding of a rigorous formative assessment routine within an innovative programme. What is more salient here is the basis, in that programme, of a model of the development of performance linked to a criterion based scheme of diagnostic assessment.

In example number *five* (Butler, 1988), the work was grounded more narrowly in an explicit psychological theory, in this case about a link between intrinsic motivation and the type of evaluation that students have been taught to expect. The experiment involved 48 11-year-old Israeli students selected from 12 classes across 4 schools, half of those selected being in the top quartile of their class on tests of mathematics and language, the other half being in the bottom quartile. The students were given two types of task in pairs, not curriculum related, one of each pair testing

convergent thinking, the other divergent. They were given written tasks to be tackled individually under supervision, with an oral introduction and explanation. Three sessions were held, with the same pair of tasks used in the first and third. Each student received one of three types of written feedback with returned work, both on the first session's work before the second, and on the second session's work before the third. The second and third sessions, including all of the receipt and reflection on the feedback, occurred on the same day. For feedback, one-third of the group were given individually composed comments on the match, or not, of their work with the criteria which had been explained to all beforehand. A second group were given only grades, derived from the scores on the preceding session's work. The third group were given both grades and comments. Scores on the work done in each of the three sessions served as outcome measures. For the 'comments only' group the scores increased by about one-third between the first and second sessions, for both types of task, and remained at this higher level for the third session. The 'comments with grade' group showed a significant decline in scores across the three sessions, particularly on the convergent task, whilst the 'grade only' group declined on both tasks between the first and last sessions, but showed a gain on the second session, in the convergent task, which was not subsequently maintained. Tests of pupils' interest also showed a similar pattern: however, the only significant difference between the high and the low achieving groups was that interest was undermined for the low achievers by either of the regimes involving feedback of grades, whereas high achievers in all three feedback groups maintained a high level of interest.

The results were discussed by the authors in terms of cognitive evaluation theory. A significant feature here is that even if feedback comments are operationally helpful for a student's work, their effect can be undermined by the negative motivational effects of the normative feedback, i.e. by giving grades. The results are consistent with literature which indicates that task-involving evaluation is more effective than ego-involving evaluation, to the extent that even the giving of praise can have a negative effect with low-achievers. They also support the view that pre-occupation with grade attainment can lower the quality of task performance, particularly on divergent tasks.

This study carries two significant messages for this general review. The first is that, whilst the experiment lacks ecological validity because it was not part of or related to normal curriculum work and was not carried out by the students' usual teachers, it nevertheless might illustrate some important lessons about ways in which formative evaluation feedback might be made more or less effective in normal classroom work. The second lesson is the possibility that, in normal classroom work, the effectiveness of formative feedback will depend upon several detailed features of its quality, and not on its mere existence or absence. A third message is that close attention needs to be given to the differential effects between low and high achievers, of any type of feedback.

The *sixth* example is in several ways similar to the fifth. In this work (Schunk, 1996), 44 students in one USA elementary school, all 9 or 10 years of age, worked over seven days on seven packages of instructional materials on fractions under the

instructions of graduate students. Students worked in four separate groups subject to different treatments—for two groups the instructors stressed learning goals (learn how to solve problems) whilst for the other two they stressed performance goals (merely solve them). For each set of goals, one group had to evaluate their problem-solving capabilities at the end of each of the first sessions, whereas the other was asked instead to complete an attitude questionnaire about the work. Outcome measures of skill, motivation and self-efficacy showed that the group given performance goals without self-evaluation came out lower than the other three on all measures. The interpretation of this result suggested that the effect of the frequent self-evaluation had out-weighed the differential effect of the two types of goal. This was confirmed in a second study in which all students undertook the self-evaluation, but on only one occasion near the end rather than after all of the first six sessions. There were two groups who differed only in the types of goal that were emphasised—the aim being to allow the goal effects to show without the possible overwhelming effect of the frequent self-evaluation. As expected, the learning goal orientation led to higher motivation and achievement outcomes than did the performance goal.

The work in this study was curriculum related, and the instructions given in all four 'treatments' were of types that might have been given by different teachers, although the high frequency of the self-evaluation sessions would be very unusual. Thus, this study comes closer to ecological validity but is nevertheless an experiment contrived outside normal class conditions. It shares with the previous (fifth) study the focus on goal orientation, but shows that this feature interacts with evaluative feedback, both within the two types of task, and whether or not the feedback is derived from an external source or from self-evaluation.

The *seventh* example involved work to develop an inquiry-based middle school science-based curriculum (Frederiksen & White, 1997). The teaching course was focused on a practical inquiry approach to learning about force and motion, and the work involved 12 classes of 30 students each in two schools. Each class was taught to a carefully constructed curriculum plan in which a sequence of conceptually based issues was explored through experiments and computer simulation, using an inquiry cycle model that was made explicit to the students. All of the work was carried out in peer groups. Each class was divided into two halves: a control group used some periods of time for a general discussion of the module, whilst an experimental group spent the same time on discussion, structured to promote reflective assessment, with both peer assessment of presentations to the class and self-assessment. This experimental work was structured around students' use of tools of systematic and reasoned inquiry, and the social context of writing and other communication modes. All students were given the same basic skills test at the outset. The outcome measures were of three types: one a mean score on projects throughout the course, one a score on two chosen projects which each student carried out independently, and one a score on a conceptual physics test. On the mean project scores, the experimental group showed a significant overall gain; however, when the students were divided into three groups according to low, medium or high scores on the initial basic skills test, the low scoring group showed

a superiority, over their control group peers, of more than three standard deviations, the medium group just over two, and the high group just over one. A similar pattern, of superiority of the experimental group which was more marked for low scoring students on the basic skills test, was also found for the other two outcomes. Amongst the students in the experimental group, those who showed the best understanding of the assessment process achieved the highest scores.

This science project again shows a version of formative assessment which is an intrinsic component of a more thorough-going innovation to change teaching and learning. Whilst the experimental-control difference here lay only in the development of 'reflective assessment' amongst the students, this work was embedded in an environment where such assessment was an intrinsic component. Two other distinctive features of this study are first, the use of outcome measures of different types, but all directly reflecting the aims of the teaching, and second the differential gains between students who would have been labelled 'low ability' and 'high ability' respectively.

The *eighth* and final example is different from the others, in that it was a meta-analysis of 21 different studies, of children ranging from pre-school to grade 12, which between them yielded 96 different effect sizes (Fuchs & Fuchs, 1986). The main focus was on work for children with mild handicaps, and on the use of the feedback to and by teachers. The studies were carefully selected—all involved comparison between experimental and control groups, and all involved assessment activities with frequencies of between 2 and 5 times per week. The mean effect size obtained was 0.70. Some of the studies also included children without handicap: these gave a mean effect size of 0.63 over 22 sets of results (not significantly different from the mean of 0.73 for the handicapped groups). The authors noted that in about half of the studies teachers worked to set rules about reviews of the data and actions to follow, whereas in the others actions were left to teachers' judgments. The former produced a mean effect size of 0.92 compared with 0.42 for the latter. Similarly, those studies in which teachers undertook to produce graphs of the progress of individual children as a guide and stimulus to action reported larger mean gains than those where this was not done (mean effect size 0.70 compared with 0.26).

Three features of this last example are of particular interest here. The first is that the authors compare the striking success of the formative approach with the unsatisfactory outcomes of programmes which had attempted to work from a priori prescriptions for individualised learning programmes for children, based on particular learning theories and diagnostic pre-tests. Such programmes embodied a deductive approach in contrast with the inductive approach of formative feedback programmes. The second feature is that the main learning gains from the formative work were only achieved when teachers were constrained to use the data in systematic ways which were new to them. The third feature is that such accumulation of evidence should have given some general impetus to the development of formative assessment—yet this paper appears to have been overlooked in most of the later literature.

Some General Issues

The studies chosen thus far are all based on quantitative comparisons of learning gains, six of them, and those reviewed in the eighth, being rigorous in using pre- and post-tests and comparison of experimental with control groups. We do not imply that useful information and insights about the topic cannot be obtained by work in other paradigms.

As mentioned in the Introduction, the ecological validity of studies is clearly important in determining the applicability of the results to normal classroom work. However, we shall assume that, given this caution, useful lessons can be learnt from studies which lie at various points between the 'normal' classroom and the special conditions set up by researchers. In this respect all of the studies exhibit some degree of movement away from 'normal' classrooms. The study (by Whiting *et al.*, 1995) which is most clearly one of normal teaching within the everyday classroom is, inevitably, the one for which quantitative comparison with a strictly equivalent control was not possible. More generally, caution must be exercised for any studies where those teaching any experimental groups are not the same teachers as those for any control groups.

Given these reservations, however, it is possible to summarise some general features which these examples illustrate and which will serve as a framework for later sections of this article.

- It is hard to see how any innovation in formative assessment can be treated as a marginal change in classroom work. All such work involves some degree of feedback between those taught and the teacher, and this is entailed in the quality of their interactions which is at the heart of pedagogy. The nature of these interactions between teachers and students, and of students with one another, will be key determinants for the outcomes of any changes, but it is difficult to obtain data about this quality from many of the published reports. The examples do exhibit part of the variety of ways in which enhanced formative work can be embedded in new modes of pedagogy. In particular, it can be a salient and explicit feature of an innovation, or an adjunct to some different and larger scale movement—such as mastery learning. In both cases it might be difficult to separate out the particular contribution of the formative feedback to any learning gains. Another evaluation problem that arises here is that almost all innovations are bound to be pursuing innovations in ends as well as in means, so that the demand for unambiguous quantitative comparisons of effectiveness can never be fully satisfied.
- Underlying the various approaches are assumptions about the psychology of learning. These can be explicit and fundamental, as in the constructivist basis of the first and the last of the examples, or in the diagnostic approach of Bergan *et al.* (1991) or implicit and pragmatic, as in the mastery learning approaches.
- For assessment to be formative the feedback information has to be used—which means that a significant aspect of any approach will be the differential treatments which are incorporated in response to the feedback. Here again assumptions

about learning, and about the structure and nature of learning tasks which will provide the best challenges for improved learning, will be significant. The different varieties and priorities across these assumptions create the possibility of a wide range of experiments involving formative assessment.

- The role of students in assessment is an important aspect, hidden because it is taken for granted in some reports, but explicit in others, particularly where self and peer assessments by and between students are an important feature (with some arguing that it is an inescapable feature—see Sadler, 1989).
- The effectiveness of formative work depends not only on the content of the feedback and associated learning opportunities, but also on the broader context of assumptions about the motivations and self-perceptions of students within which it occurs. In particular, feedback which is directed to the objective needs revealed, with the assumption that each student can and will succeed, has a very different effect from that feedback which is subjective in mentioning comparison with peers, with the assumption—albeit covert—that some students are not as able as others and so cannot expect full success.

However, the consistent feature across the variety of these examples is that they all show that attention to formative assessment can lead to significant learning gains. Although there is no guarantee that it will do so irrespective of the context and the particular approach adopted, we have not come across any report of negative effects following on an enhancement of formative practice. In this respect, one general message of the Crooks review has been further supported.

One example, the kindergarten study of Bergan *et al.* (1991) brings out dramatically the importance that may be attached to the achievement of such gains. This particular innovation has changed the life chances of many children. This sharp reality may not look as important as it really is when a result is presented dryly in terms of effect sizes of (say) 0.4 standard deviations.

To glean more from the published work, it is necessary to change gear and move away from holistic descriptions of selected examples to a more analytic form of presentation. This will be undertaken in the next five sections.

Assessment by Teachers

Current Practice

Teachers' practices in formative assessment were reviewed in the articles by Crooks (1988) and Black (1993b). Several common features emerged from these surveys. The overall picture was one of weak practice. Key weaknesses were:

- Classroom evaluation practices generally encourage superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge which pupils soon forget.
- Teachers do not generally review the assessment questions that they use and do not discuss them critically with peers, so there is little reflection on what is being assessed.

- The grading function is over-emphasised and the learning function under-emphasised.
- There is a tendency to use a normative rather than a criterion approach, which emphasises competition between pupils rather than personal improvement of each. 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.

More recent research has confirmed this general picture. Teachers appear to be unaware of the assessment work of colleagues and do not trust or use their assessment results (Cizek *et al.*, 1995; Hall *et al.*, 1997). Both in questioning and written work, teachers' assessment focuses on low-level aims, mainly recall. There is little focus on such outcomes as speculation and critical reflection (Stiggins *et al.*, 1989; Schilling *et al.*, 1990; Pijl, 1992; Bol & Strage, 1996; Senk *et al.*, 1997), and students focus on getting through the tasks and resist attempts to engage in risky cognitive activities (Duschl & Gitomer, 1997). Although teachers can predict the performance of their pupils on external tests—albeit tests reflecting low-level aims—their own assessments do not tell them what they need to know about their students' learning (Lorsbach *et al.*, 1992; Rudman, 1987).

Reviews of primary school practices in England and in Greece have reported that teachers' records tend to emphasise the quantity of students' work rather than its quality, and that whilst tasks are often framed in cognitive terms, the assessments are in affective terms, with emphasis on social and managerial functions (Bennett *et al.*, 1992; Pollard *et al.*, 1994; Mavromattis, 1996). There are some striking comments by those who have researched these issues—one report on science practices sees formative and diagnostic assessment as 'being seriously in need of development' (Russell *et al.*, 1995, p. 489), another closes with a puzzled question 'Why is the extent and nature of formative assessment in science so impoverished?' (Daws & Singh, 1996, p. 99), whilst a survey of teachers in Quebec Province, Canada, reports that for formative assessment 'Indeed they pay lip service to it but consider that its practice is unrealistic in the present educational context' (quoted by Dassa *et al.*, 1993, p. 116). The conclusion of a survey about practice in Belgian primary schools was that the criteria used by teachers were 'virtually invalid by external standards' (Grisay, 1991, p. 104). A study which used interviews and so produced a richer picture of the perceptions of US teachers concludes as follows:

> Most of the teachers in this study were caught in conflicts among belief systems, and institutional structures, agendas, and values. The point of friction among these conflicts was assessment, which was associated with very powerful feelings of being overwhelmed, and of insecurity, guilt, frustration, and anger. These teachers expressed difficulty in keeping track of and having the language to talk about children's literate development. They also described pressure from external accountability testing. They differed in their assessment strategies and in the language they used to describe students' literacy development. Those who worked in highly controlling situations were inclined to use blaming language and tended to

> provide global, negative descriptive assessments in impersonal language. Their assessments were likely to be based on a simple, linear notion of literacy. The less controlling the situation the less this was likely to occur. This study suggests that assessment, as it occurs in schools, is far from a merely technical problem. Rather, it is deeply social and personal. (Johnston *et al.*, 1995, p. 359)

This last quotation also draws attention to the dominance of external summative testing. The effects here run deep, witness the evidence in Britain that when teachers were required to undertake their own assessments they imitated the external tests (Bennett *et al.*, 1992), and seemed to be able to think only in terms of frequent summative tests with no feedback action (Ratcliffe, 1992; Harlen & Malcolm, 1996). A similar effect was encountered in assessment reforms in Queensland (Butler & Beasley, 1987). A different tension between formative and summative assessment arises when teachers are responsible for both functions: there has been debate between those who draw attention to the difficulties of combining the two roles (Simpson, 1990; Scott, 1991; Harlen *et al.*, 1992) and those who argue that it can be done and indeed must be done to escape the dominance of external summative testing (Black, 1993a; Wiliam & Black, 1996). The requirement in Scotland, that teachers use external tests when they think their pupils are ready, and mainly for moderation purposes (i.e. checking for consistency of standards between schools), does not seem to have resolved these tensions (Harlen *et al.*, 1995).

Assessment, Pedagogy and Innovation

Given these problems, it is not surprising that when national or local assessment policies are changed, teachers become confused. Several of the reports quoted above give evidence of this. A patchy implementation is reported for reforms of teacher assessment in France (Broadfoot *et al.*, 1996) and in French Canada (Dassa, 1990), whilst in the UK such changes have produced a diversity of practices, some of which may be counter-productive and in conflict with the stated aims of the changes which triggered them (McCallum *et al.*, 1993; Gipps *et al.*, 1997). Where changes have been introduced with substantial training or as an intrinsic part of a project in which teachers have been closely involved, the pace of change is slow because it is very difficult for teachers to change practices which are closely embedded within their whole pattern of pedagogy (Torrie, 1989; Shepard *et al.*, 1994, 1996; Shepard, 1995) and many lack the interpretive frameworks that they need to co-ordinate the many separate bits of assessment information in the light of broad learning purposes (Bachor & Anderson, 1994). Indeed, some such work fails to produce its effect. A project with teachers in the creative arts, which tried to train them to communicate with students in order to appreciate the students' view of their own work, found that despite the training, many teachers stuck to their own agenda and failed to respond to cues or clues from the students which could have re-oriented that agenda (Radnor, 1994).

The issue that emerges here, as it did in the section above on Classroom

experience, is the close link of formative assessment practice both with other components of a teacher's own pedagogy, and with a teacher's conception of his or her role. In a project aimed at enhancing the power of science teachers to observe their students at work, teachers could not find time for observing because they were not prepared to change classroom practices in order to give students more free responsibility and give themselves a less closely demanding control. The authors interpreted this as a reluctance to break the existing symbiosis of mutual dependency between teachers and students (Cavendish *et al.*, 1990). In research with special education teachers, Allinder (1995) found that teachers with a strong belief in their high personal and teaching efficacy made better use of formative assessment than their less confident peers.

We have not tried here to give a comprehensive review of the literature on teachers' assessment practices. The aim has been to highlight some key points which are relevant to the main purpose of this review. The three outstanding features are:

- that formative assessment is not well understood by teachers and is weak in practice;
- that the context of national or local requirements for certification and accountability will exert a powerful influence on its practice; and
- that its implementation calls for rather deep changes both in teachers' perceptions of their own role in relation to their students and in their classroom practice.

These features have implications for research into this area. Research which simply interrogates existing practice can probably do little more than confirm the rather discouraging findings reported above. To be productive therefore, research has to be linked with a programme of intervention. If such intervention is to seek implementation with and through teachers in their normal classrooms, it will be changing their roles and ways of teaching; then the formative initiative will be part of a larger pattern of changes and its evaluation must be seen in that larger context. More closely focused pieces of research might be more attractive as ways of exploring the different issues that are involved, but might have to use imported researchers because teachers cannot be expected quickly to abandon habitual roles and methods for a limited experiment. Thus at least some of the research that is needed will inevitably lack ecological validity.

Students and Formative Assessment

The core of the activity of formative assessment lies in the sequence of two actions. The first is the perception by the learner of a gap between a desired goal and his or her present state (of knowledge, and/or understanding, and/or skill). The second is the action taken by the learner to close that gap in order to attain the desired goal (Ramaprasad, 1983; Sadler, 1989). For the first action, the prime responsibility for generating the information may lie with the student in self-assessment, or with another person, notably the teacher, who discerns and interprets the gap and communicates a message about it to the student. Whatever the procedures by which the assessment message is generated, in relation to action taken by the learner it

would be a mistake to regard the student as the passive recipient of a call to action. There are complex links between the way in which the message is received, the way in which that perception motivates a selection amongst different courses of action, and the learning activity which may or may not follow.

For the purposes of this review, the involvement of students in formative assessment will be considered by division into two broad topics, as follows:

(1) The first of these will focus on those factors which influence the reception of the message and the personal decisions about how to respond to it. The concern will be with the effects of beliefs about the goals of learning, about one's capacity to respond, about the risks involved in responding in various ways, and about what learning work should be like: all of these affect the motivation to take action, the selection of a line of action and the nature of one's commitment to it.
(2) The second will focus on the different ways in which positive action may be taken and the regimes and working contexts in which that action may be carried out. The focus here will be on study methods, study skills, collaboration with peers, and on the possibilities of peer and self-assessment.

There is clearly a strong interaction between the two areas. In particular, if self and peer-assessment are promoted in a classroom, this affects the initial generation of the message about a gap as well as the way in which a learner may work to close it. However, the over-arching sets of beliefs to be considered within the first focus bear on the perception of and response to feedback messages, albeit in different ways, whether they are generated by the self or by others. In the studies reported within the first topic, both sources of feedback have been considered.

Reception and Response

In his analysis of formative assessment by teachers in France, Perrenoud comments that:

> A number of pupils do not aspire to learn as much as possible, but are content to 'get by', to get through the period, the day or the year without any major disaster, having made time for activities other than school work [...] Formative assessment invariably presupposes a shift in this equilibrium point towards more school work, a more serious attitude to learning [...] Every teacher who wants to practise formative assessment *must reconstruct the teaching contracts so as to counteract the habits acquired by his pupils.* Moreover, some of the children and adolescents with whom he is dealing are imprisoned in the identity of a bad pupil and an opponent. (Perrenoud, 1991, p. 92 (author's italics))

This rather pessimistic view is supported, but modified, by the finding of Swain (1991) that some secondary students working on teacher assessed science projects in England would respond to serious difficulties by working on subsidiary aspects of the task, so avoiding the main problem, and would be 'insatiable' in their search for

cues for the 'right answer' from teachers. These symptoms of insecurity were accompanied by frequent moves to secure the esteem of the teacher. Similarly, Blumenfeld reports (1992) that some US students will try to avoid the risks involved in tackling a challenging assignment.

Thus whilst reluctance to be drawn into a more serious engagement with learning work may arise from a wish merely to minimise effort, there can be other influences. One problem may be fear—the extra personal commitment required can carry with it an enhanced penalty for failure in terms of one's self-esteem. Another problem may be that students can fail to recognise formative feedback as a helpful signal and guide (Tunstall & Gipps, 1996a). Purdie & Hattie's (1996) comparative study of the responses of Japanese and Australian students, which aimed to explore their self-regulation strategies, shows that response can be culturally determined. Many researches report that positive learning gains secured by formative feedback are associated with more positive attitudes to learning—notably in mastery learning regimes where the use to be made of the feedback is clearly planned (Kulik *et al.*, 1990; Whiting *et al.*, 1995), but there can also be negative affects and the notions of attitude and motivation have to be explored in more detail if the origin of such effects is to be understood.

In the review and analysis presented by Blumenfeld (1992), he points to evidence that students can be reluctant to seek help, and are not always happy to receive extra assistance because it is interpreted as evidence of their low ability. Similarly, in their experimental study of the effects of different forms of guidance with 3rd and 6th graders solving mathematical problems, Newman & Schwager (1995) found that, whilst the different approaches could make a difference, the frequency of requests for help from all students was surprisingly low and they concluded that there is a need to encourage more help-seeking in the ordinary classroom. The central feature of this particular study was that the difference between the two forms of feedback guidance being given was a seemingly narrow one. One group were told that the goals of the work were in *learning* ('This will help you to learn new things...') with emphasis on the importance of understanding how to tackle problems of the type presented, whilst for the other the goal stressed was their own *performance* ('How you do helps us to know how smart you are and what kind of grade you will get...') with corresponding emphasis on completing as many problems as possible. Apart from this difference, all received the same tuition, including feedback, in respect of the work and all were encouraged to seek for help whenever they felt the need. The *performance* goal students were more likely to show maladaptive questioning patterns and solved fewer problems, particularly when those initially classified as low achievers were compared across the two groups.

Goal Orientation

This effect of goal orientation on learning has been extensively studied. The study of Ames & Archer (1988) involved only enquiry into the goals that students already held. They found that their sample of 176 students ranging over grades 8 to 11 could be divided into two groups—those with mastery orientation and those with

performance orientation. The former spoke of the importance of learning, believed in the value of effort to achieve mastery, and had a generally positive attitude to learning. The latter attributed failure to lack of ability, spoke more in terms of their relative ability, about learning with relatively little effort if able, and focused on the significance of out-performing others. A similar distinction was made in the intervention study by Butler (1988) already described in the section on Classroom experience above in which the terms 'ego-involving feedback' and 'task-involving feedback' were used. The surprising result of this study, that the giving of grades could undermine the positive help given by task comments, illustrates the sensitivity of the issues raised here. In a later study, Butler & Neuman (1995) showed that those in task mode were more likely to seek help and to explain help-avoidance in terms of seeking independent mastery, whilst those in an ego mode sought help less and explained their avoidance in terms of masking their incapacity. Two general reviews of this field both stress that feedback which draws attention away from the task and towards self-esteem can have a negative effect on attitudes and performance (Cameron & Pierce, 1994; Kluger & DeNisi, 1996). It is even the case that giving praise can have bad effects, particularly when it is not linked to objective feedback about the work. Lepper & Hodell (1989) argue that reward systems can undermine both interest and motivation, whilst a detailed study by Pryor & Torrance (1996) shows how a teacher can concentrate on protective care for a child at the expense of helping the child to learn.

Several studies by Schunk (Schunk, 1996) have developed this same theme. This has already been brought out in the one described in the section on Classroom experience. In two studies, one on the learning of reading with 5th grade remedial students (Schunk & Rice, 1991), the other on writing instruction with mainstream 5th-graders (Schunk & Swartz, 1993a), the second showed that better results were secured by giving process goals rather than product goals, and both showed that where the feedback on process goals was supplemented to include information about students' progress towards the overall aim of the learning, both the students' learning performance and their beliefs about their own performance capacities (self-efficacy), were at the highest level. The patterns of association between achievement, self concept, and the regimes of study and feedback experienced by students have been the subject of a detailed analysis, using results from 12 high school biology courses, by Thomas *et al.* (1993). A complex pattern of links emerged, but the importance of self-concept was clear, and it also seemed that the provision of challenging assignments and extensive feedback lead to greater student engagement and higher achievement.

Self-perception

In a more general review of the literature in this field, Ames (1992) started from the evidence about the advantages that 'mastery' (i.e. task-related) goals can secure and reviews the salient features of the learning environments that can help to secure these advantages. She concludes that evaluation to students should focus on individual improvement and mastery, but before this the tasks proposed should help

students to establish their own self-referenced goals by offering a meaningful, interesting and reasonably demanding challenge. She also recommends that feedback should be private, must be linked to opportunities for improvement, and should encourage the view that mistakes are a part of learning. The self-perception of students is all-important here, and this will be strongly influenced by teachers' beliefs about the relative importance of 'effort' as against 'ability' in their views of learning. In particular, it is important that motivation is seen to involve changes in students' qualitative beliefs about themselves, which the setting of goals and the style of feedback should both be designed to secure. The use of extrinsic rewards can be counter-productive if they focus attention on 'ability' rather than on the belief that one's effort can produce success. Of course, the beliefs of peers and of parents can also affect the ways in which the self-concepts of students are developed, as is pointed out in Blumenfeld's analysis (1992), which draws general conclusions similar to those of Ames.

There is evidence from many studies that learners' beliefs about their own capacity as learners can affect their achievement. Examples that can be added to those already quoted above are those of Lan *et al.* (1994), Craven *et al.* (1991), Fernandes & Fontana (1996), King (1994) and Butler & Winne (1995). The study of Fernandes & Fontana showed that achievements within the experiment in Portugal described in the section on Classroom experience were linked to an enhancement of the students' sense of their own control over their learning, and King's work also focused on locus of control as a predictor of performance. Grolnick & Ryan (1987) demonstrated that self-directed learning styles produced better conceptual learning, an effect that they attributed to enhanced autonomy and internal locus of control. These issues were analysed in a theoretical paper by Deci & Ryan (1994) which is discussed further in the section on Meta-task processes.

Studies by Skaalvik (1990), Siero & van Oudenhoven (1995) and Vispoel & Austin (1995) all show that the reasons students gave for the results of their learning differ between low achievers, who attribute failure to low ability, and high achievers who tend to attribute success to effort. Vispoel & Austin urge that teachers should help students to overcome attributions to ability, and should encourage them to regard ability as a collection of skills that they can master over time.

Craven's work in mathematics and reading with students in grades 3 to 6 (Craven *et al.*, 1991), showed that students' self-concept could be enhanced by feedback designed to this end and that whilst those whose self-concept was initially low showed large gains, those with initially high self-concept showed no gains. In addition, the students' attribution of success in the work to effort increased whilst attributions to ability did not. However, in this short intervention, the results obtained by the researcher could not be replicated by the teacher and there were no significant differences in achievement between experiment and control groups. A final and further perspective is added by the review of Butler & Winne (1995), who, in addition to covering the evidence that many of the factors mentioned above can have on learning achievement, also draw attention to the importance of learners' beliefs about the importance of effort, about the amount of effort that successful learning can demand, about the nature of learning, and about the—immature—

expectation that all learning should lead to simple and unambiguous answers to all the questions that can be raised.

Overall, this section of this review has been selective and does not claim to cover the many possible aspects implied in the terms attitude and motivation. The particular focus in the work reviewed here is to call attention to the importance of a variety of personal features—self-concept, self-attribution, self-efficacy, and assumptions about the nature of learning. There are clearly complex overlaps and interactions between these features; Geisler-Brenstein & Schmeck (1995) in a comprehensive analysis of evidence on these inter-relationships, have formulated an 'Inventory of Learning Processes' in order to promote what they call 'a multi-faceted perspective on individual differences in learning'.

The importance of these features arises from the conjunction of two types of research results summarised above. One is that the 'personal features' referred to above can have important effects on a student's learning. The other is that the way in which formative information is conveyed to a student, and the context of classroom culture and beliefs about ability and effort within which feedback is interpreted by the individual recipient, can affect these personal features for good or ill. The hopeful message is that innovations which have paid careful attention to these features have produced significant learning gains when compared with the existing norms of classroom practice.

Assessment by Students

The focus of this section is to discuss one aspect of the learning activity which may follow from the student acceptance and understanding of the need to close a gap between present achievement and desirable goals. In formative assessment, any teacher has a choice between two options. The first is to aim to develop the capacity of the student to recognise and appraise any gaps and leave to the student the responsibility for planning and carrying out any remedial action that may be needed. This first option implies the development within students of the capacity to assess themselves, and perhaps to collaborate in assessing one another. The second option is for teachers to take responsibility themselves for generating the stimulus information and directing the activity which follows. The first of these two will be the subject of this section, whilst the second will be discussed in the sections titled Strategies and tactics for teachers and Systems below. The two options overlap in that it is possible to combine the two approaches: the boundary between this section and the section on Strategies and tactics for teachers will therefore be arbitrary, as is the boundary between this section and the section on Classroom experience.

The focus on self-assessment by students is not common practice, even amongst those teachers who take assessment seriously. Daws & Singh (1996) found that only about a third of the UK science teachers whom they sampled involved pupils directly in their own assessment in any way, and both Parkin & Richards (in Fairbrother *et al.*, 1994, pp. 15–28) and the account of Norwegian initiative by Jernquist (reported in Black & Atkin, 1996, pp. 92–119) describe the introduction of self-assessment,

respectively in secondary school science in the UK and in secondary mathematics in Norway, as innovations. In the general literature on classroom assessment, the topic is frequently overlooked—for example, the otherwise comprehensive collection by Phye (1997) contains no piece which focuses explicitly on self- and peer-assessment.

The motives for introducing this practice are diverse. Parkin & Richards started because of the practical impossibility of appraising the level of need of each individual in a class of about 30 students engaged in practical laboratory work—if they could do it for themselves the teacher could deploy his/her effort more efficiently. In his review of the literature on student self-evaluation in professional training courses in the health sciences, Arthur (1995) reported that the requisite skills are not purposefully taught in most programmes, but also described new research to develop these skills in nursing education. The motive given here is that the future professional will need all of the skills necessary for life-long learning, and self-evaluation must be one of these.

The Norwegian initiative started from a more fundamental motive, which was to see self- and peer-assessment as an intrinsic part of any programme which aims to help students to take more responsibility for their own learning. A different slant on this aspect is provided in the study by James of recorded dialogues between teachers and students (1990). This study showed that in such dialogues, the teacher's power easily overwhelms the student's contribution, the latter being too modestly tentative. The effect is that inquiry into the reasons for a student's difficulty is not pursued. Some of the research discussed in the section on Classroom experience above involved experiments where work on goals was pursued both with and without training in self-evaluation; an example is the research by Schunk (1996) which showed that, if combined with performance goals, self-evaluation practice improved persistence, self-efficacy and achievement.

Some authors have taken the argument further by developing a theoretical reflection on how students might change their understandings. The assumption here is they cannot do so unless they can first understand the goals which they are failing to attain, develop at the same time an overview in which they can locate their own position in relation to those goals, and then proceed to pursue and internalise learning which changes their understanding (Sadler, 1989). In this view, self-assessment is a *sine qua non* for effective learning. This theoretical stance will be further explored at the end of this section and in the section titled Prospects for the theory and practice of formative assessment.

Studies of Self-assessment

Research studies of self- and peer-assessment can be broadly divided into two categories—those involving experimental work yielding quantitative data on achievement and those for which the evidence is qualitative. These will now be discussed in turn. Two quantitative examples have already been described in some detail in the section on Classroom experience (Fontana & Fernandes, 1994; Frederiksen & White, 1997). Both of these have in common an emphasis on the need for students to understand the learning goals, to understand the assessment criteria, and to have

the opportunity to reflect on their work. Peer evaluation played a part only in the Frederiksen & White study.

Two studies have worked with children who have learning difficulties. In the first of these (McCurdy & Shapiro, 1992), the oral reading rates of elementary school students were improved by giving them verbal and visual performance feedback, either by the teacher only, or through peer-monitoring, or self-monitoring. The largest gains, measured by comparison of pre- and post-test scores over the programme's period of nine weeks, were achieved by the self-monitoring group, whilst all three did better than a control group who had no formative feedback. Both on the grounds of acceptability to the teachers involved and on the reliability of their own appraisal of their work, the peer- and self-monitoring methods were preferred and one benefit of both was that they reduced the amount of time that the special education teachers had to spend on measurement in their classrooms. In the second research (Sawyer *et al.*, 1992) the focus was on the writing composition skills of 4th and 5th grade students. Here, a group who were taught self-regulated strategies with explicit attention to goals did better than a similar group without the goal emphasis and a group without self-monitoring instruction. The first group were better overall on generalisation of the writing skills taught, but all groups with feedback did better, after the particular experiment was over, than other learning disability students without any experience of such feedback.

In research to investigate the most effective way of using a problem-solving software programme (Delclos & Harrington, 1991), two groups of 5th and 6th grade students were both given training in their pro-active use of the programme, but one of them also had to take part in monitoring exercises, described by the authors as meta-cognitive training. There was also a matched control group who used the programme without the training. The monitoring exercises were provided by a booklet of questions with which students monitored their results on a set of practice problem-solving exercises selected from the software. Both trained groups achieved greater success with the programme than the control group, but those with the monitoring training were also significantly better than those without it. They were more successful with the more complex problems, they succeeded more quickly, and overall they were seen to be employing more effective strategies. They seemed to do better, not because they could use the particular strategies more effectively, but because they started by reflecting on a problem and considering the possibilities of using different strategies before proceeding—an outcome which seemed to link with the meta-cognitive emphasis underlying the self-monitoring training.

A focus on self-directed learning was seen, in the review by Thomas (1993), to be a necessary concomitant to the moves to develop practical work, study skills, and responsibility for learning amongst students. He distinguished course features that discourage independent learning, such as test review handouts, from those that encourage it, including extensive performance feedback, and reviewed evidence which established that such activities can improve student achievement. In a review of the practice of writing, Zimmerman & Risemberg (1997) discussed the different forms of the practice of self-regulation employed by several well-known authors and linked this to research evidence on the effectiveness of supporting students by

encouraging self-monitoring (Schunk & Swartz, 1993b; Zimmerman & Bamdura, 1994). A closely related set of studies by King (1994) on students' questioning strategies will be reviewed in the section on Questions below.

Self-evaluation is an intrinsic aspect of reflection on one's own learning. Several qualitative studies report on innovations designed to foster such self-reflection. In science education, Baird *et al.* (1991) reported on work with 27 teachers and 350 students where teachers were helped to know more about their students and to learn more about how they might change the style of classroom work by a strategy based on meta-cognition and constructivism. Both the teachers and the students involved had to analyse what had happened in a piece of the learning work, and each side had to propose three changes to be put into effect. Later, students had to evaluate whether these changes had happened. The evidence, based on self-reports by those involved, was that successful implementations had been achieved. Maqsud & Pillai (1991) trained a class of high-school students in self-scoring of their tests and found that their score gains were significantly higher than those of a control group class: they attributed this to the lowering of their students' normal distrust of and antagonism towards marked feedback. Similar success was achieved by Merrett & Merrett (1992) in an experiment aimed to help students to realise, through feedback on their self-assessment, the lack of correspondence between their self-perception of their work and the judgments of others; the quality and depth of the students' self-assessments were enhanced as the experiment proceeded. Similar work is reported by Griffiths & Davies (1993), Powell & Makin (1994) and Meyer & Woodruff (1997).

A larger scale innovation is fully described in a book by Ross *et al.* (1993). The aim was to change assessment of achievement in the visual arts by bringing students into the assessment process as reflective practitioners, mainly through the development of 'assessment conversations' in which students were encouraged to reflect on their work and to articulate their reflections. The authors are enthusiastic in their accounts of the success of their work, and believe that the students involved showed that they 'are capable of rich and sophisticated responses to and understandings of their own work ... in collaboration with their conversation partner' (p. 161). They concluded that the approach opened up new opportunities in aesthetic knowing and appraisal, but that it also required that teachers abandon traditional assessment practices. However, the evidence of the 'success' of the work is to be found only in the accounts, illustrated with quotations, of the quality of the students' aesthetic judgments. Similarly qualitative reports were given of an initiative to hand over all responsibility for assessment of a first-year undergraduate course to students' self-assessment (Edwards & Sutton, 1991), and of the outcome of a project to train 2nd, 3rd, and 4th grade students to record their on or off task state of work at regular intervals (Wheldall & Pangagopolou-Steamatelatou, 1992). In both cases, the initiative produced a significant change in students' commitment to their work and there was also some indirect evidence in both of improvement in their learning achievement.

Peer-assessment

Several of the accounts described in this section involve both self-assessment and peer-assessment. Peer-assessment as such is included in several accounts of the development of group collaboration as a part of classroom learning activity. In an experimental study by Koch & Shulamith (1991), college students were taught to generate their own questions about topics in physics, and achieved better learning gains than those who used only teacher's questions; amongst those generating their own questions, some also used peer feedback to answer and discuss their efforts, and this group showed even greater learning gains than the rest. Higgins *et al.* (1994) also used collaborative work, in their work with 1st and 2nd grade school-children developing assessment skills in their integrated project work. The children generated their own criteria, and the quality of these rose during the study. Good agreement with teachers' assessments was achieved, with children tending to under-assess. However, groups were not accurate in their assessments of other groups. The reliabilities of self- and peer-assessments were also investigated, in work with college biology students, by Stefani (1994). He found correlations with teachers' assessments of 0.71 for self-assessments and 0.89 for peer-assessments. All of the students said that the self- and peer-assessment work made them think more, and 85% said that it made them learn more. Hughes & Large (1993) also investigated peer-assessment of final year undergraduates in pharmacology and found a correlation coefficient of 0.83 between the mean ratings of peers and those of a group of staff.

It is often difficult to disentangle the peer-assessment activity from other novel activities in work of this kind, and impossible in general to ascribe any reported gains to the assessment component. General reviews are given by Slavin (1991) and by Webb (1995). The second of these does focus on assessment practices in group work and it stresses the importance of training in group processes and of the setting of clear goals and clear achievement criteria. In such groups, a clear choice has to be made, and shared in the group, between a goal of the best performance from the group as a group, and a goal of improving individuals' performances through group collaboration. The question of the optimum group composition is a complex one; where a group goal has priority, then for well defined tasks, established high achievers are the most productive, but for more open tasks a range of types of students is an advantage. Where individuals' performance has priority, then the high achievers are little affected by the mix, but the low achievers benefit more from a mixed group provided that the group training emphasises methods for drawing out, rather than overwhelming, their contribution. The need for such care is emphasised in a study of group discussions in science education by Solomon (1991).

Links to Theories of Learning

The arguments given by Zessoules & Gardner (1991) show how any assessment changes of the types described above might be expected to enhance learning if they help students to develop reflective habits of mind. They further argue that such

development should be an essential component in programmes for the implementation of authentic assessment in classroom practice. Assessment is to be seen as a moment of learning, and students have to be active in their own assessment and to picture their own learning in the light of an understanding of what it means to get better.

In summary, it can be seen that these various approaches to developing self-assessment by pupils hold promise of success. However, their interpretation in relation to more general theories of learning raises fundamental problems, as illustrated by the analysis of Tittle (1994). Full discussion of this and similar work will be deferred until the last part of this article. A few points can be introduced here. In a review of European research in this field, Elshout-Mohr (1994) points out both that students are often unwilling to give up misunderstandings—they need to be convinced through discussion which promotes their own reflection on their thinking—and also that if a student cannot plan and carry out systematic remedial learning work for himself, he or she will not be able to make use of good formative feedback. Both of these indicate that self-assessment is essential. Similarly, Hattie *et al.* (1996) argue that direct teaching of study skills to students without attention to reflective, meta-cognitive, development may well be pointless. One reason for the need to look for radical change is that students bring to their work models of learning which may be an obstacle to their own learning. That pupils do have such models that are to a degree culturally determined is illustrated by a comparison of the approaches to learning of Australian and Japanese students (Purdie & Hattie, 1996), whilst the finding that the most able students in either country are more alike than their peers in having developed similar effective habits of learning shows that such constraining traditions can be overcome.

The task of developing students' self-assessment capabilities may be approached as a task of providing them with appropriate models of this way of working. In a modest way, Carroll (1994) tried to do this by providing worked examples of algebra problems to students for them to study, replacing some of the work on solving problems for themselves which they would normally be doing. The achievement of these students was improved by this method, and the low achievers showed particularly good improvements. The author proposed that for many students, the task of tackling new problems in a new area of work might not be useful because of cognitive overload. Study of a worked example provided a less loaded learning situation in which reflection on the processes used could be developed. More generally, Bonniol's discussion (1991) leads to the conclusion that the teacher must provide a model of problem-solving for the student, and needs also to be able to understand the model in the head of the learner so that he/she can help the learner to bring order into his or her 'meta-cognitive haze'. The difficulty here is that many teachers do not have a good model of problem-solving and of effective reasoning to transmit, and therefore lack both the theoretical framework within which to interpret the evidence provided by students and the model to which to direct them in the development of their own self-assessment criteria.

Strategies and Tactics for Teachers

Overview

The various aspects of a teacher's work in formative assessment can be organised in relation to the temporal sequence of decisions and actions that are entailed. This approach will be used here as a framework for describing the work reported in the literature. Thus, the sub-sections below will deal in turn with the choice of tasks, with classroom discourse, with several aspects of the use of questions, with tests and then with feedback from tests. A closing section will then look at strategies overall, including work that looks more deeply at the assumptions and rationales that might underlie the articulation of tactics.

Choice of Task

It is obvious that formative assessment which guides learners towards valued learning goals can only be generated with tasks that both work to those goals and that are open in their structure to the generation and display of relevant evidence, both from student to teacher and to students themselves. In their detailed qualitative study of the classroom characteristics of two outstandingly successful high-school science teachers, Garnett & Tobin (1989) concluded that the key to their success was the way they were able to monitor for understanding. A common feature was the diversity of class activities—with an emphasis on frequent questioning in which 60% of the questions were asked by the students. In a more general review of classroom environments, Ames (1992) selects three main features which characterise successful 'mastery' (as opposed to 'performance'—see section on Goal orientation above) classrooms. The first of these is the nature of the tasks set, which should be novel and varied in interest, offer reasonable challenge, help students develop short-term self-referenced goals, focus on meaningful aspects of learning and support the development and use of effective learning strategies. Blumenfeld (1992) explores some of these issues, pointing out that such notions as 'challenging' and 'meaningful' are problematic. A task where the challenge goes too far can lead to student avoidance of the risks involved, and for students who are far behind it is difficult to encourage their efforts without at the same time making them aware of how far behind they are. Similarly, tasks can be meaningful for a variety of reasons and it is important to emphasise those meanings which might be productive for learning.

In an earlier review about science teaching, Dumas-Carre & Larcher (1987) were more ambitious. They emphasised the need to shift current pedagogy to give more emphasis to procedural aspects of knowledge and less to the declarative aspects. They outlined a scheme for the comparative analysis of tasks which could be deployed by teachers to produce a descriptive analysis of the tasks they were using. This scheme distinguished tasks which (a) presented a specific situation identical to the one studied, or (b) presented a 'typical' problem but not one identical to the one studied, requiring identification of the appropriate algorithm and its use, rather than exact replication of an earlier procedure as in (a), and (c) a quite new problem

requiring new reasoning and construction of a new approach, deploying established knowledge in a new way. Students would need special and explicit training for tackling tasks of type (c). They recommended that all three types of task are needed, but that teachers do not currently plan or analyse the tasks that they set by any scheme of this type. Such foresight is an essential condition for planning the incorporation of formative assessment, both for provision of feedback and for planning how to respond to it.

Discourse

That the quality of the discourse between teacher and students can be analysed at several different levels is evident from the extensive literature on classroom interactions and discourse analysis. In a review of questioning in classrooms, Carlsen (1991) contrasts the process–product approach with the socio-linguistic paradigm, and argues that inconsistency of research results on the cognitive level of questioning may be due to neglect of the fact that the meaning of a question cannot be inferred from its surface features alone. As both he and Filer (1995) argue, the meaning behind any discourse depends also on the context, on the ways in which questions in a particular classroom have come to signify the patterns of relationships between those involved that have been built up over time. Pryor & Torrance (1996) give an example of how a habitual pattern can be unhelpful for learning. Newmann (1992) makes a plea, on similar grounds, for assessment in social studies to focus on discourse, defined by him as language produced by the student with the intention of giving narrative, argument, explanation or analysis. The plea is based on an argument that current methods, in which students are constrained to use the language of others, undermine the constructive use of discourse and so trivialise social knowledge. A striking example of such effects is reported in a paper by Filer (1993). In the same vein, Quicke & Winter (1994) report success in work with low achieving students in Year 8, where they aimed to develop a social framework for dialogue about learning. The work of Ross *et al.* (1993) in aesthetic assessment in the arts can be seen as a response to this plea, and the difficulties reported by Radnor (1994) are evidence of the inadequacy of established practice.

Radnor's paper uses the phrase 'qualitative formative assessment' in its title, and this may help to explain why quantitative evidence for the learning effects of discourse variations is hard to find. An exception is the research by Clarke (1988) on classroom dialogue in science classrooms. He analysed the discourse of three teachers in four classrooms, grading the quality of the discourse by summation over four criteria. These included the numbers of interpretable themes, the numbers of cross-correlations (an indicator of coherence) and proportions of themes explicitly related to the content of the lessons. This discourse variable was included with three other measures, of scholastic aptitude, locus of control and Piagetian level respectively, as independent variables, and an achievement post-test as dependent variable. With the class as the unit of analysis, the discourse variable accounted for 63% of

the variance, with the three others accounting respectively for under 4%, 22% and 14%.

Johnson & Johnson (1990) present a meta-analysis to show that collaborative discourse can produce significant gains in learning. Rodrigues & Bell (1995), Cosgrove & Schaverien (1996) and Duschl & Gitomer (1997) all report work with science teachers to promote such discourse. The concern in all three cases was to help students to move, in talk about their work, from a focus in everyday and content based terms towards a deeper discussion of conceptual learning. Roth & Roychoudhury (1994) recommend the use of concept maps as an aid in such discussions; such maps, drawn by the students, serve to provide useful points of reference in clarifying the points under discussion and enable the teacher to engage in 'dynamic assessment'.

Questions

Some of the relevant aspects of questioning by teachers have already been introduced above. The quality of classroom questioning is a matter for concern, as expressed in the work of Stiggins *et al.* (1989) who studied 36 teachers over a range of subjects and over grades 2 to 12, by observation of classroom work, study of their documentation, and interviews. At all levels the questioning was dominated by recall questions, and whilst those trained to teach higher-order thinking skills asked more relevant questions, their use of higher-order questions was still infrequent. An example of the overall result was that in science classrooms, 65% of the questions were for recall, with only 17% on inferential and deductive reasoning. The patterns for written work were similar to those for the oral work. Bromme & Steinberg (1994) studied the classroom strategies of novice teachers in mathematics and found that they tended to treat students' questions as being from individual learners, whereas the responses of expert teachers tended to be directed more to a 'collective student'.

Several authors report work focused on question generation by students, and as pointed out in the section on Self-assessment above, this may be seen as an extension of work on students' self-assessment. With college students, King (1990, 1992a,b; 1994) found that training which prompted students to generate specific thought-provoking questions and then attempt to answer them is more effective than training in other study techniques, which she interprets in terms of the strategy underlying the training which aimed to develop learner autonomy and learners' control over their own work. Similar results, which also showed that students' own questions produced better results than adjunct questions from the teacher, were reported for Israeli college students by Koch & Shulamith (1991). In work with 5th grade school students, a similar approach was used in training students with problem-solving on computer administered tasks (King, 1991). With a sample of 46 students, one group was given no extra instruction, another was trained to ask and answer questions with student partners, whilst a third group were also trained in questioning one another in pairs but directed to use strategic questions for guidance in cognitive and meta-cognitive activity. The latter training focused on the use of 'generic questions' such as 'How are X and Y alike?' and 'What would happen if...?'.

The outcome was measured by a post-test of written problems and a novel computer task. The group trained to ask strategic questions of one another out-performed the others. Foos *et al.* (1994) have reported similar success on outcome measures when students were trained to prepare for examinations by several techniques, the most successful being the generation of their own study questions followed by attempts to answer them. This work, and similar work in school science classes (King & Rosenshine, 1993) can be seen as part of a larger strategy to promote inquiry-based critical thinking (King, 1995).

Such work has two main elements. One is the promotion of higher-order thinking and self-regulation of their study by students through the generation of questions, the other is to conduct this development through peer interaction. A comprehensive review of studies of this type by Rosenshine and colleagues (1996) presents a meta-analysis of selected studies. The effects are strongly positive, but the effect size depends on whether the outcome measure is a standardised test, or a comprehension test developed by the experimenter. The latter give larger effects, with means of 1.00 for 5 studies with reciprocal peer questioning and 0.88 for 11 others without this feature (the difference between these was not significant). The conclusion is that there is no evidence that peer interaction is superior to direct instruction in question generation. It is pointed out that the theoretical rationale for active processing by students does not provide specific guidance about the choice of methods, and the review discusses the diverse approaches adopted in some detail.

A rather different use of questioning is to explore and develop students' prior knowledge. A review of work of this type (Pressley *et al.*, 1992) establishes that requiring learners to compose answers with explanations to explore their prior knowledge of new work does improve learning, and that this may be because it helps the learner to relate the new to the old and to avoid superficial judgments about the new content.

Another category of questioning is the use of adjunct questions with text. There is little to add here to the studies reviewed by Crooks. A study by Holliday & Benson (1991) with a high school biology class showed that when the teacher required work on questions of the type to be used in a test and emphasised their importance, performance was improved. Another study with the use of comprehension adjunct questions was aimed to improve concept learning with science lessons in which computer-animated sequences were used (Holliday & McGuire, 1992). Eighth grade students were assigned to a control group or to one of four treatment groups. The results confirmed that the questions used did succeed in their aim of focusing students' attention on the concepts involved, and that where they were used for only the first 8 of the 12 sequences used, they produced effects on the way in which those remaining were studied. The authors composed the questions to serve a general aim of scaffolding the meta-cognitive activity of the students. However, such work should be appraised in the light of the results quoted in the section on Choice of task above (and Otero & Campanario, 1990; Carroll, 1994) which indicated that it might help to employ some of the learning time of students on other critical tasks in place of the time spent tackling questions.

The Use of Tests

One study giving evidence that frequent testing can lead to improved learning has already been cited in the section on Classroom experience above (Martinez & Martinez, 1992). Bangert-Drowns, *et al.* (1991b) reviewed the evidence of the effects of frequent class testing. Their meta-analysis of 40 relevant studies showed that performance improved with frequent testing and increased with increased frequency up to a certain level, but that beyond that (somewhere beyond 1 and 2 tests per week) it could decline again. The evidence also indicated that several short tests were more effective than fewer longer ones. Similar evidence is quoted by Dempster (1991, 1992, see the section on Task motivation processes, below).

In a later investigation with college students in psychology, Iverson *et al.* (1994) found that the addition of frequent ungraded tests produced no significant improvement in performance, even though the students in the experiment said that they would like to have such tests in other courses also. A similarly negative result was reported by Strawitz (1989), but a contrary, positive effect was found by Schloss *et al.* (1990) working with graduate students in teacher training for special education. When given a short formative quiz after each lecture students performed significantly better than they did when no quiz was provided on three measures—post-tests of familiar items, post-tests of unfamiliar items and a survey of satisfaction with the instruction.

In some subject areas, teachers are reluctant to use tests for fear of inhibiting creativity. Gilbert (1996) attempted to tackle this problem with primary school teachers in the assessment of art. The project worked with both experienced and trainee teachers to develop a framework and a language for assessing art, and from this the group were able to formulate guidance about the feedback appropriate to children according to the classification in the framework to which their work was judged to correspond. Developments of new assessment methods appropriate to specific subjects are also described by Adelman *et al.* (1990) for visual and performance arts with older secondary students and by Harnett (1993) for history in primary schools.

The aim and structure of the tests employed are not explained in most studies. Crooks speculated that low level aims might benefit from more frequent testing, but that higher level aims would benefit from a lower frequency. Khalaf & Hanna (1992) conducted a more searching review and disagreed with Crooks on this point. They selected 20 studies, of which 18 gave positive effects. They point out that the nature of the criterion test could distort results. In any such study, the criterion test might be more important to a control group than to the treatment group. On the other hand, if the final summative test were to contain questions similar to those in the class tests, there could be a distortion in the reverse direction. They concluded that only four of the studies they reviewed were free from this type of flaw. The results of these were all positive with a mean effect size of 0.37, but they were all with college students. Following their critical review, these authors describe the results of an investigation with 2000 students in 93, 10th grade, classes in Saudi Arabia. Control classes were given the normal monthly quizzes, whilst the others were given

bi-monthly quizzes. The criteria were tests set by an investigator who had not seen any of the quizzes constructed and used by the teachers, and included a test at the end of the course and a delayed test three months later. Treatment and control classes were set up in pairs having the same teacher for the two members of each pair. There were significant differences between the two groups on both test occasions, with an effect size of about 0.3 in favour of the more frequently tested group. The effect was much greater for high achievers than for the medium and low achievers. However, the tests used were composed only of true-false or multiple choice items, so that the learning at issue was relatively superficial in nature. The care taken with this experiment shows how studies cannot be accepted as relevant without careful scrutiny of the experiment design, and of the quality of the questions used for both the treatment and the criterion tests.

Behind these reservations lies the larger issue of whether or not the testing is serving the formative assessment function. This cannot be clarified without a study of how test results are interpreted by the students. If the tests are not used to give feedback about learning, and if they are no more than indicators of a final high-stakes summative test, or if they are components of a continuous assessment scheme so that they all bear a high-stakes implication, then the situation can amount to no more than frequent summative testing. Tan (1992) describes a situation in a course for first year medical students in which he collected evidence that the frequent summative tests were having a profound negative influence on their learning. The tests called only for low-level skills and had thereby established a 'hidden curriculum' which inhibited high level conceptual development and which meant that students were not being taught to apply theory to practice. Similar concerns about the cognitive level of testing work is expressed by Hall *et al.* (1995, further discussed in the section titled Expectations and the social setting, below).

The Quality of Feedback

Both of the previous sub-sections lead to the almost obvious point that the quality of the feedback provided is a key feature in any procedure for formative assessment. The instructional effect of feedback from tests was reviewed by Bangert-Drowns *et al.* (1991a) using a meta-analysis of 58 experiments taken from 40 reports. Effects of feedback were reduced if students had access to the answers before the feedback was conveyed. When this effect had been allowed for, it was then the quality of the feedback which was the largest influence on performance. Programmed instruction and simple completion assessment items were associated with the smallest effects. Feedback was most effective when it was designed to stimulate correction of errors through a thoughtful approach to them in relation to the original learning relevant to the task.

The feedback provided by teachers' written responses to students' homework was studied in an experiment with over 500 Venezuelan students involving 18 mathematics teachers in three schools (Elawar & Corno, 1985). They trained the teachers to give written feedback which concentrated on specific errors and on poor strategy, with suggestions about how to improve, the whole being guided by a focus on deep

rather than superficial learning. A control group followed the normal practice of marking the homework without comments. In order to check whether effects of the feedback training on their teaching could account for any results, a third group of the trained teachers marked half of their classes with full feedback and the other half with marks only. All were given a pre-test and one of three parallel forms of post-test. Analysis of variance of the results showed a big effect associated with the feedback treatment, which accounted for 24% of the variance in final achievement (with another 24% associated with prior achievement). The treatment also reduced the initial superiority of boys over girls and had a large positive effect on attitudes towards mathematics.

In a quite different sphere of learning, Tenenbaum & Goldring (1989) carried out a meta-analysis with 16 studies of the effects of 'enhanced instruction', involving emphasis on cues, participation, reinforcement, feedback and correctives, on motor skill learning in physical education. Exposure to these forms of enhancement produced gains with a mean effect size of 0.66, and they also enhanced students' time on task.

The linkage of feedback to assumptions about the nature of the student learning which it is designed to encourage has been taken further in work on curriculum-based assessment by Fuchs *et al.* (1991). Their experiment with mathematics students explored the possible enrichment of a scheme of systematic assessment of student development by setting up an 'expert system' which teachers could consult to guide their instructional planning in relation to the students' assessment results. The experiment used three groups of teachers, one who used no systematic assessment, a second group who used such assessment, and a third who used the same assessment together with the expert system. Both of the second and third groups revised their teaching programmes more frequently than the first. However, only the third group produced better student achievement than the first, and whilst teachers in the second group responded to feedback by using different problems without changing the teaching strategies, those in the third reviewed both. The conclusion reached was that teachers need more than good assessment instruments—they also need help to develop methods to interpret and respond to the results in a formative way. One requirement for such an approach is a sound model of students' progression in the learning of the subject matter, so that the criteria that guide the formative strategy can be matched to students' trajectories in learning; this need, and some evidence bearing on ways to meet it, has been studied for both school mathematics and school science (Black, 1993a, pp. 58–61; Masters & Evans, 1986; Brown & Denvir, 1987). For such data, the criterion sequences have to be attuned by normative data to indicate reasonable expectations for students at different ages. This has been attempted with data for spelling, reading and mathematics by Fuchs *et al.* (1993), who consider both the practical needs and the implications of such data for developmental studies of academic progress.

A more comprehensive discussion of feedback will be offered in a section below devoted to this topic.

Formulation of Strategy

The sub-sections above can be regarded as treatments of the various components of a kit of parts that can be assembled to compose a complete strategy. The research studies described can be judged valuable, in that they explore a complex situation by treating one variable at a time, or as flawed because any one tactic will vary in its effect with the holistic context within which it operates. It has also emerged that at least some of the accounts are incomplete in that the quality of the procedures or instruments that evoke the feedback, and of the assumptions which inform the interpretation of that feedback, cannot be judged. At the same time it is clear that the fundamental assumptions about learning on which the procedures, instruments and interpretations are based are all important.

Several authors have written about the larger strategic picture, and reference has already been made to some of their arguments. The analysis of Thomas *et al.* (1993) stands out because they have attempted a quantitative study to encompass the many variables involved. They applied hierarchical linear modelling to data collected from 12 high school biology courses, focusing on the features of the courses which placed demands on and gave the support to the students. At the student level, the results indicated a positive link between achievement and both their self-concept of academic ability and their study activities; these last two were also linked with one another. Students' engagement in active study work was positively associated with the provision of challenging activities and with extensive feedback on their work in the course. Such feedback was also directly related with high achievement. Amongst the other relationships that were teased out was the finding that instructor support which reduced course demands strengthened the relationship between self-concept and achievement. This work constitutes an ambitious effort, but, almost of necessity, only very general information is provided about the underlying quality of the work being studied.

Weston *et al.* (1995) have argued that if the literature on formative assessment is to inform instructional design, then a common language is needed. They identify four components—who participates, what roles can be taken, what techniques can be used and in what situations these can occur, and argue that instructional design should be based on explicit decisions about these four, to be taken in the light of the goals of the instruction. The model was used to analyse 11 instructional tests and revealed that there were many assumptions about these four issues which were embedded in the language about formative evaluation.

Both Ames and Nichols attempt more ambitiously detailed analyses. For Ames (1992), the distinction between the performance and mastery perspectives is a starting point, but she then outlines three salient features, namely meaningful tasks, the promotion of the learners' independence by giving authority to their own decision making, and evaluation which focuses on individual improvement and mastery. The importance of changing the assumptions that teachers make about learning is recognised in this review. The analysis bears many similarities to that of Zessoules & Gardner (1991). An account of a project to provoke and support teachers in making changes of this type (Torrie, 1989) brings

out the many difficulties that teachers encountered, both in making their assessments relating to learning criteria, and in changing their teaching and feedback to break away from norm-referenced assumptions in supporting student's learning.

Nichols's (1994) analysis goes deeper in concentrating on what he terms cognitively diagnostic assessments. It is pointed out that classical psychometrics has been directed towards the use of assessments to guide selection, and so a new relationship with cognitive science is needed if it is to be used to guide learning. Tests must be designed in the light of models of specific knowledge structures in order to help determine the progress of learners in acquiring those structures, so that the interpretation of the feedback can serve the purpose of making inferences about students' cognitive mechanisms. It is clear that many traditional types of test are inadequate to this purpose because they do not reveal the methods used by those tested. Lorsbach *et al.* (1992) explored the factors which affect the validity of assessment tasks when judged from a constructivist perspective and emphasised that a major threat to validity is the extent to which students can construct the meanings of the tasks intended by those who set them. Both in their account and that of Torrance & Pryor (1995) the analysis is illustrated by detailed records of the work of one or two teachers. However, the latter authors, developing the arguments in Torrance (1993), provide a more wide-ranging theoretical discussion, contrasting two approaches to formative assessment—a behaviourist one, stressing measurement against objectives, and a social constructivist one integrating the assessment into learning. Similarly, for the assessment of language, Shohamy (1995) argues that the complexity of language calls for a special discipline for language assessment, grounded in a clear theoretical perspective of what it means to 'know a language'.

Thus task selection, and the type of feedback that a task might generate, require a cognitive theory which can inform the link between learners' understanding and their interactions with assessment tasks, in the light of which assessment activities can be designed and interpreted. Such an approach will of course interact strongly with the pedagogy adopted, and may have to temper an a priori theoretical position with a readiness to adapt and develop by an inductive approach as formative feedback challenges the rationale of the work (Fuchs & Fuchs, 1986). The conclusion of the analysis is that a very substantial effort, involving collaboration between psychometricians, cognitive scientists and subject experts is needed.

All of these discussions point to the need for very far-reaching changes if formative evaluation is to realise its potential. Some large-scale changes in pedagogy have attempted to meet such targets, and are distinguished from what has been discussed in this section by their comprehensive and strategic approach. These will be the subject of the next section.

Systems

General Strategies

Good assessment feedback is either explicitly mentioned or strongly implied in

reports of a range of studies and initiatives in which such feedback is one component of a broader strategy. Thus, for example, in summarising a study of school effectiveness Mortimore *et al.* (1988) point out that feedback and good record-keeping are key aspects of effectiveness. In the initiative in Britain to develop a holistic 'Record of Achievement' to cover all aspects of a student's work and contribution within a school, the student is to be involved in negotiating an agreed record. Thus self-assessment is reported to be an important feature, but it has been incorporated in a variety of ways and is sometimes superficial (Broadfoot, 1992; Broadfoot *et al.*, 1990). Enhanced attention to diagnosis and remediation is a feature of many other schemes, for example reading recovery and Slavin's Success for All scheme (Slavin *et al.*, 1992, 1996).

Assessment and feedback are also an important feature of mastery learning programmes, discussed in more detail below, but with many (if not all) of these teaching systems, even identifying the precise nature of the formative feedback used, let alone its contribution to the global improvements in attainment generated, is difficult. For this reason, these systems are reviewed only briefly in what follows.

Studies of Mastery Learning

Mastery learning originated as a practical implementation of the learning theories of John B. Carroll. He proposed that success in learning was a function solely of the ratio of the time actually spent learning to the time needed for learning—in other words, any student could learn anything if they studied it long enough. The time spent learning depended both on the time allowed for learning and the learner's perseverance while the time needed to learn depended on the learner's aptitude, the quality of the teaching, and the learner's ability to understand this teaching (see Block & Burns, 1976, p. 6). Two main approaches to mastery learning were developed in the 1960s. One, developed by Benjamin Bloom, using teacher-paced group-based teaching approaches, was called Learning for Mastery (LFM) and the other was Keller's individual-based, student-paced Personalized System of Instruction (PSI). The vast majority of the research undertaken into mastery learning has been centred on Bloom's, rather than Keller's model, and that which has been done on PSI is largely confined to further and higher education.

A key consequence of Bloom's LFM model is that students of differing aptitude will differ in their achievements unless those with less aptitude are given either a greater opportunity to learn or better quality teaching. For most proponents of mastery learning, this is not to be achieved by targeting teaching resources at students of lower aptitude, but by improving the quality of teaching for all students, the underlying assumption being that students with higher aptitude are better able to make sense of incomplete or poor instruction (Milkent & Roth, 1989).

The key elements in this strategy, according to McNeil (1969) were:

- The learner must understand the nature of the task to be learned and the procedure to be followed in learning it.
- The specific instructional objectives relating to the learning task must be formulated.

- It is useful to break a course or subject into small units of learning and to test at the end of each unit.
- The teacher should provide feedback about each learner's particular errors and difficulties after each test.
- The teacher must find ways to alter the time some students have available to learn.
- It may be profitable to provide alternative learning opportunities.
- Student effort is increased when small groups of two or three students meet regularly for as long as an hour to review their test results and to help one another overcome the difficulties identified by means of the test.

Therefore, although the principles of mastery learning cover all aspects of learning and teaching, effective formative assessment is a key component of effective mastery learning. Three major reviews of the research into the effectiveness of mastery learning use the technique of 'meta-analysis' to combine the results from a variety of different studies. The review of Block & Burns (1976) covers work done in the first half of the 1970s while Guskey & Gates (1986) and Kulik *et al.* (1990) cover the subsequent decade.

Between them, the reviews by Block & Burns (1976) and Guskey & Gates (1986) provide 83 measures (from 35 studies) of the effect of mastery learning on general achievement, all using the 'Learning For Mastery' approach (LFM). They found an average effect size of 0.82, which is equivalent to raising the achievement of an 'average' student to that of the top 20%, and one of the largest average effects ever reported for a teaching strategy (Kulik & Kulik, 1989). When the age of the students involved is examined, it appears as if mastery learning is less effective for older students. However it is not clear whether this is because older students are more 'set in their ways' and therefore have more difficulty in changing their ways of working to those required for mastery learning, or because mastery learning is adapted more readily within primary school curricula and pedagogy.

These reviews have also considered whether mastery learning is more effective in some subjects than others. Block & Burns (1976) found that results for science (and according to some authors, sociology) were more consistent, but lower than for other subjects, while results for mathematics were, on average, higher, but far less consistent. However, while Guskey & Gates (1986) also found low effect sizes for science, these were comparable to the effect sizes for mathematics, and both were substantially lower than those found for language arts and social studies. Thus no clear consensus emerges from research on the relative effectiveness of mastery learning programmes in different subjects.

The 1990 review by Kulik *et al.* looked at 108 studies which were judged to meet their criteria for inclusion in a meta-analysis. Of these, 91 were carried out with students over 18 years of age, 72 using Keller's PSI approach and 19 using Bloom's LFM approach. The 17 school-based studies all used LFM, although these were also skewed towards older students—only two of the studies contained any results from students younger than 11 years of age. The effect sizes found were smaller than those found by Block & Burns and Guskey & Gates—not surprising given the greater representation of studies with older students. However, Kulik *et al.* also

found that the self-paced PSI approach tended to have smaller effect sizes than the teacher-paced LFM approach, and also seems to reduce completion rates in college courses. The three reviews also found that mastery-learning programmes are more effective for lower-achieving students, thus tending (as originally intended) to reduce the range of achievement in the cohort, although others, such as Livingstone & Gentile (1996), have found no evidence to support the 'decreasing variability hypothesis'.

Parallel to the issue of whether mastery learning is more effective for lower-achieving students is that of whether it is equally effective for all teachers. Martinez & Martinez (1992) looked at the effect of repeated testing in a remedial undergraduate mathematics course and found that frequent 'mastery' testing was effective in raising achievement, but it was more effective for the less experienced teacher. Based on a meta-analysis of 40 studies, Bangert & Drowns *et al.* (1991b) estimated that testing once every three weeks showed an effect-size of 0.5 over no testing, increasing to around 0.6 for weekly tests and 0.75 for twice-weekly tests.

Others, most notably Robert Slavin, have questioned whether mastery learning is effective at all. In his own review of research on mastery learning, he criticises meta-analysis as being too crude, because of the way that results from all the research studies that satisfy the inclusion criteria are averaged. His own approach is to use a 'best-evidence' synthesis (Slavin, 1987), attaching more (necessarily subjective) weight to the studies that are well-designed and conducted. Although many of his findings agree with the meta-analytic reviews described above, he points out that almost all of the large effect sizes have been found on teacher-prepared, rather than standardised, tests, and indeed, the effect sizes for mastery learning measured by standardised tests are close to zero. This suggests that the effectiveness of mastery learning might depend on the 'curriculum-embeddedness' of the outcome measures. This is supported by Kulik *et al.*'s (1990) finding that the effect sizes for mastery learning as measured by formative tests (typically around 1.17) are greater than for summative tests (around 0.6).

Slavin (1987) argues that this is because, in mastery learning studies where outcome is measured using teacher-produced tests, the teachers focus narrowly on the content that will be tested. In other words, the effects are produced (either consciously or unconsciously) by 'teaching to the test'. The crux of this disagreement is therefore the measure of 'mastery' of a domain—should it be the teacher-produced test or the standardised test?

The Relevance of Mastery Learning

The only clear messages emerging from the mastery learning literature are that mastery learning appears to be effective in raising students' scores on teacher-produced tests, is more effective in teacher-paced programmes than in self-paced programmes, and is more effective for younger students.

However, while establishing that under certain circumstances, mastery learning is effective in raising achievement, the literature gives very little evidence as to which aspects of mastery learning programmes are effective. For example, while

most of the research concentrates on the effects of mastery learning on students, the explanation of the effects could be that preparing for teaching for mastery provides professional development for the teacher (Whiting *et al.*, 1995).

Indeed, one of the criticisms voiced by the reviewers cited above is that too often it is impossible to establish from the research reports which features of mastery learning were implemented in the research being reported, let alone which were effective (Guskey & Pigott, 1988). There are at least five aspects of 'typical' mastery learning programmes that are relevant to the purposes of the present review:

- that students are given feedback;
- that students are given feedback on their current achievement against some expected level of achievement (i.e. the 'mastery' level);
- that such feedback is given rapidly;
- that such feedback is (or is at least intended to be) diagnostic;
- that students are given the opportunity to discuss with their peers how to remedy any weaknesses.

However, it is by no means clear that all of these are necessary to achieving the gains claimed for mastery learning. For example, Kulik & Kulik (1987) argue that mastery testing is a crucial component in the success of mastery learning, and that its omission leads to a substantial drop in a programme's effectiveness. On the other hand, programmes without formalised 'mastery' testing, but with many of the other features of 'mastery' programmes can show significant effects.

For example, in a study by Brown *et al.* (1996) a group of low-achieving second grade students were given a year of 'transactional strategies instruction' (TSI), in which teachers explained and modelled strategies, gave additional coaching as needed, and encouraged students to explain to each other how they used the strategies. At the end of the year, this group outperformed by a considerable margin a similar group taught by highly regarded teachers using more traditional methods (exact effect sizes are not given, but they range between 1 and 2 standard deviations).

Assessment Driven Models

An account has already been given in the section on Classroom experience of the introduction of a complete system of planning and measurement for kindergarten children, in which the wholesale innovation appears to have had formative assessment as a central component, so that it seems valid to attribute the reported success to that component (Bergan *et al.*, 1991). Another wholesale approach is to reform discourse through application of the concept of scaffolding (Day & Cordon, 1993; Hogan & Pressley, 1997). Different again are approaches where the problem of acting on assessment results is tackled by constructing the work on a particular module or topic in such a way that the basic ideas have been covered by about two-thirds of the way through the course; assessment evidence is reviewed at this stage, so that in the remaining time differentiated work can proceed according to the needs of different pupils (Black & Dockrell, 1984; Dwight, 1988).

Two more loosely defined systems are those described as 'curriculum-based assessment' and those described as 'portfolio' systems.

Curriculum-based assessment (CBA) is a development which expanded in the late 1980s. The focus of many of the studies has been on early years education and the identification of pupils with special educational needs, but its methods and principles could apply right across the spectrum of education. Useful reviews of the literature are to be found in the collection edited by Kramer (1993), and one article in that collection (Shinn & Good III, 1993) sets out the central features of CBA as follows:

- Assessment exercises should faithfully reflect the main learning aims and should be designed to evoke evidence about learning needs.
- The main purpose for assessment is the formative purpose.
- Validity is paramount—seen as ensuring that instructional decisions taken on the basis of assessment evidence are justified.
- The focus of attention is the individual learner and individually attuned remedial action.
- The information from assessment should serve to locate the individual's attainment in relation to criteria for learning, but that this location should also be informed by norm data on the progress of others working to the same curriculum.
- The assessment should be frequent so that the trajectory of learning over time can be traced: the gradient of learning success is the key indicator—to follow each pupil's progress in general, and to indicate cases of special need.

Shinn does not make a sharp distinction between CBA and the related concept of Curriculum-based measurement (CBM), but others insist on this distinction (Salvia & Hughes, 1990; Salvia & Ysseldyke, 1991; Deno, 1993; Tindal, 1993). Deno, for example, sees CBM as a sub-set of CBA concerned with specific measures and procedures focused on basic skills for the diagnostic work of special education teachers, and also regards the title's term 'Measurement' as reflecting the importance in the strategy of relating measures to an established quantitative scale. There is also some disagreement as to whether or not CBA can be described as a 'behavioural' approach.

A further precision seen as important by two authors is to distinguish CBA from mastery learning (Deno, 1993; Fuchs, 1993). They see mastery learning as requiring that learners follow a specific skill sequence step by step, which constrains learning to follow a particular path, whereas CBA is far broader and looser and so allows for learners to follow a variety of routes to learn, with less emphasis on treating discrete skills in isolation.

The research evidence about CBM is reviewed by Fuchs (1993). For her, the setting of explicit learning goals is a distinctive feature of CBM. The research evidence is that students achieve higher levels of attainment if the learning goals are ambitious for them. Experiments have also compared those working to static goals, set at the outset and not subsequently amended, with those working to dynamic goals, which are amended, usually up-graded with corresponding changes in the instruction, in the light of measured progress. The dynamic approach leads to better achievements.

The description by Shinn of what he calls the paradigm of CBA makes it clear that this is a formative assessment approach, and that many of its features would be essential in any incorporation of formative assessment into a learning programme. What may be distinctive is the insistence on sharply focused test designs, and on the use of frequent tests to give graphs of performance against time as a key diagnostic instrument.

Portfolios

The portfolio movement is more closely associated with efforts to change the impact of high-stakes, often standardised, testing of school learning. There is a vast literature associated with the portfolio movement in the USA. Much of it is reviewed, by Collins (1992), in the edited collections of Belanoff & Dickson (1991) and—for assessment of writing—by Calfee & Perfumo (1996a), whilst Courts & McInerney (1993), set out some of the issues in higher education. Mills (1996) gives an account of the origins of the innovation, describing the work as an attempt in Vermont to satisfy demands of accountability whilst avoiding the pressures of standardised tests.

A portfolio is a collection of a student's work, usually constructed by selection from a larger corpus and often presented with a reflective piece written by the student to justify the selection. The involvement of the student in reviewing and selecting is seen as central—as Mills says 'Inventing ways to promote that kind of reflection on a wide scale has been at the heart of the Vermont assessment from the beginning' (Mills, 1996, p. 192) and, speaking of the response of students 'What was striking was their ability to reflect on their own work in relation to a set of internalised standards—standards that they shared with many others' (Mills, 1996, p. 194). Similarly, writing about a different, national, project, Daro (1996) reports on the innovators' enthusiasm, both for the power of portfolios to focus student attention on their own learning efforts and accomplishments, and for the evidence that teachers believe the work changes the ways in which they teach and increases their expectations for their students. Calfee & Freedman (1996) see portfolios as offering a technology for helping the slogan of 'student-centred learning' to become a reality. Others (Herman *et al.*, 1996) emphasise that it is valuable for students to understand the assessment criteria for themselves, whilst Yancey (1996), in a more subtle analysis of the concept of reflection as learning, points out that the practice of helping students to reflect on their work has made teachers more reflective for themselves.

However, there is little by way of research evidence, that goes beyond the reports of teachers, to establish the learning advantages. Attention has focused rather on the reliability of teachers' scoring of portfolios because of the motive to make them satisfy concerns for accountability, and so to serve summative purposes as well as the formative. In this regard, the tension between the purposes plays out both in the selection and in the scoring of tasks (Benoit & Yang, 1996). Daro (1996) describes scoring approaches based on a multi-dimensional approach, with the criterion that each dimension reflect an aspect of learning which can be understood by students and which reflects an important aspect of learning. However, he identifies the problem thus 'But it does not necessarily follow that it will be practical to bring national standards

into the focus of self-assessing students and their teachers' (Davo, 1996, p. 241).

Calfee & Perfumo (1996b) report on research with teachers into their experience of using portfolios. The results showed a disturbing gap between the general rhetoric and actual practice, for they showed that many teachers were paying little attention to external standards and were producing little evidence of any engagement of students in understanding why they are doing this work. Their conclusion was that the future fate of the movement hung in the balance, either between three negative possibilities—anarchy, disappearance, or becoming too standardised—or the positive possibility of promoting a profound revolution in learning.

Slater *et al.* (1997) describe an experiment in an introductory algebra course for college students which produced no significant difference in achievement between a group engaged in portfolio production and a control group. However, the achievement test was a 24-item multiple choice test, which might not have reflected some of the advantages of the portfolio approach, and at the same time the teacher reported that the portfolio group ended up asking more questions about real world applications and had been led to discuss more complex and interesting phenomena than the control group. In chapter 3 of their book, Courts & McInerney (1993) also report on an attempt to evaluate a writing course with college students which also seemed to show only small gains, but point out that they were assessing holistic writing qualities which previous assessments and learning experiences of their students had neglected.

Summative Examination Models

The Graded Assessment schemes in England were comprehensive provisions designed to replace terminal examinations for public certificates by a series of graded assessments, conducted in schools but moderated (i.e. checked for consistency of standards between schools) by an examining authority. In that they replaced the terminal examinations by frequent tests within each school, and enhanced the importance of components of assessed coursework as contributors to the summative results, they influenced the ways in which assessment was operated within schools and provided a distinctive scenario for the working out of formative–summative tensions. Whilst general accounts of these schemes have been published (Pennycuick & Murphy, 1986; Lock & Ferriman, 1988; Swain, 1988, 1989; Ferriman & Lock, 1989; Iredale, 1990; Lock & Wheatley, 1990) there does not appear to be any published research which could identify the particular developments of the formative functions within these schemes. A similar scheme in sciences, in that the summative function was linked to frequent assessment over extended periods within classroom work, has been described by Ratcliffe (1992). In all of these accounts, one of the problems that stands out is the difficulties that teachers and developers met in trying to establish a criterion-referenced approach to assessment. In several such schemes, an important feature has been the provision of a central bank of assessment questions from which teachers can draw according to their particular needs—but these have generally been designed with summative needs in mind. In Canada, Dassa *et al.* (1993) describe the setting up of a bank of diagnostic items organised in a three-dimensional scheme: diagnostic context, notional content and cognitive ability, the items being derived from

a study of common errors so that they could provide a basis for causal diagnosis. The overall purpose was to help teachers to provide formative personal feedback within the constraints of normal classrooms. Trials in five classrooms showed that by comparison with a control set of five more classrooms, those using the item bank had superior gains, the mean effect size being 0.7.

The problems of developing criterion-referenced assessment beset the far more radical reforms in Queensland (Withers, 1987; Butler, 1995). This Australian state abolished external examinations for secondary schools in 1971, but subsequently encountered problems in the quality and the norm-referencing of school-based assessments. Butler's article chronicles the development of a criterion-referenced approach, with teachers having to learn the skills and the state having to develop systems to ensure comparability of interpretation of the criterion standards. Greater emphasis on assessment spread over two years within classroom work, on feedback to students on successive assessment results, and on the production of student portfolios as evidence for the moderation procedures, have been required in these developments. However, the impact of these on the formative role of assessment has yet to be researched.

The systems described in this category should indeed have implications for formative assessment and could approach the problem of the formative–summative relationship from a different direction from most other studies, where the high-stakes pressures are either ignored, or accepted in that achievements on the existing measures are used (problematically) as the criteria of success. However, there is little evidence that the formative–summative relationship has been thought out in their design, and little substantial evidence about how it has worked out in practice (but see Rowe & Hill, 1996 and section titled Are there implications for policy? below).

Feedback

The two concepts of formative assessment and of feedback overlap strongly. The term feedback has occurred frequently in the account so far, and the section on the quality of feedback is explicitly concerned with the feedback function. However, that section had a limited focus, and the usages generally have been diverse and not subject to stringent consistency. Because of its centrality in formative assessment, it is important to explore and clarify the concept. This will be done in this section as a necessary prologue to the fuller review of formative assessment in the subsequent final section.

The Nature of Feedback

Originally, feedback was used to describe an arrangement in electrical and electronic circuits whereby information about the level of an 'output' signal (specifically the gap between the actual level of the output signal and some defined 'reference' level) was fed back into one of the system's inputs. Where the effect of this was to reduce the gap, it was called negative feedback, and where the effect of the feedback was to increase the gap, it was called 'positive feedback'.

In applying this model to the behavioural sciences, we can identify four elements making up the feedback system:

- data on the actual level of some measurable attribute;
- data on the reference level of that attribute;
- a mechanism for comparing the two levels, and generating information about the gap between the two levels;
- a mechanism by which the information can be used to alter the gap.

For Kluger & DeNisi (1996) only the first of these is necessary for feedback to exist. They define 'feedback interventions' as 'actions taken by an external agent to provide information regarding some aspects of one's task performance', although it is worth noting that the requirement for an external agent excludes self-regulation. In contrast, Ramaprasad (1983) defines feedback as follows:

> Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way (p. 4).

and specifically requires that for feedback to exist, the information about the gap must be used to alter the gap. If the information is not actually used in altering the gap, then there is no feedback.

For the purposes of this review, we have taken a broad view of what constitutes feedback, rather than exclude important evidence.

One of the most important reviews of the effectiveness of feedback was carried out by Kluger & DeNisi (1996). They reviewed over 3000 reports of the effects of feedback on performance (2500 papers and 500 technical reports). After excluding those without adequate controls, those where the feedback interventions were confounded with other effects, where fewer than 10 participants were included in the study, where performance was only discussed rather than measured, and those where insufficient details were given to estimate effect sizes, they were left with 131 reports, yielding 607 effect sizes, and involving 12,652 participants.

They found an average effect size of 0.4 (equivalent to raising the achievement of the average student to the 65th percentile), but the standard deviation of the effect sizes was almost 1, and around two in every five effects were negative. The fact that so many research reports found that feedback can have negative effects on performance suggests that these are not merely artefacts of poor design, or unreliability in the measures, but real, substantive effects.

In order to explain the variability in the reported effect sizes, they examined possible 'moderators' of the effectiveness of feedback interventions—that is factors which impact, either negatively or positively, on the effectiveness of feedback.

They began by noting that presented with a 'gap' between actual and reference levels of some attribute (what Kluger & DeNisi, 1996, term a 'feedback-standard discrepancy'), there are four broad classes of action.

The first is to attempt to reach the standard or reference level, which is the typical response when the goal is clear, where the individual has a high commitment to achieving the goal and where the individual's belief in eventual success is high. The second type of response is to abandon the standard completely, which is particularly

common where the individual's belief in eventual success is low (leading to 'learned helplessness'—Dweck, 1986). A third, and less extreme, response is to change the standard, rather than abandoning it altogether. Individuals may lower the standard, especially likely where they cannot or do not want to abandon it, and conversely, may, if successful, choose to raise the standard. The fourth response to a feedback-standard gap is simply to deny it exists.

Kluger & DeNisi (1996) found empirical support for each of these categories of response, and developed a theoretical model that accounted for a significant proportion of the variability in effect sizes found in the literature. They identified three levels of linked processes involved in the regulation of task performance: *meta-task processes*, involving the self; *task-motivation processes*, involving the focal task; and *task learning processes* involving the *details* of the focal task.

Meta-task Processes

In proposing a typology of teacher feedback, based on classroom research, Tunstall & Gipps (1996b) arranged their various types in a spectrum, ranging from those that direct attention to the task and to learning methods, to those which direct attention to the self—in the extreme forms by stress only on rewards and punishments. These authors did not study effects on learning, but such studies by others (e.g. Siero & van Oudenhoven, 1995) show that feedback interventions that cue individuals to direct attention to the self rather than the task appear to be likely to have negative effects on performance. Thus praise, like other cues which draw attention to self-esteem and away from the task, generally has a *negative* effect (and goes some way to explaining why several studies, such as Good & Grouws (1975), found that the most effective teachers actually praise *less* than average).

This may explain the results obtained by Boulet *et al.* (1990). A group of 80 Canadian students in their third year of secondary schooling were randomly assigned to one of three groups for a course on the writing of the major scales in music (there were no differences between the groups in terms of musical aptitude, previous academic success or ability to learn). During the course of their instruction, the first experimental group (GE1) were given feedback on their pre-test in the form of written praise, a list of weaknesses and a workplan for further instruction, while the second experimental group (GE2) were given oral feedback, told about their errors and given the opportunity to correct them. On the post-test, the second experimental group had gained more than either the first experimental group or the control group (which were not significantly different). One interpretation of this result is that the oral delivery of feedback is more effective than written delivery of feedback. However, it seems more plausible that the congratulatory message which prefaced the written feedback cued the pupils into a focus on meta-task processes, rather than on the tasks themselves.

Further evidence of the negative effect of cueing pupils to focus on the self rather than the task comes from a study carried out by Butler (1987) in which she examined the effects of four kinds of feedback (comments, grades, praise, no feedback) on the performance of 200 Israeli grade 5 and 6 students in divergent thinking tasks. Although the four groups were matched on pre-test scores, the students given comments scored

		Value system	
		External	Internal
Locus of motivation	External	External regulation	Identified regulation
	Internal	Introjected regulation	Integrated regulation

FIG. 1. Classification of behaviour regulation, based on Deci & Ryan (1994).

one standard deviation higher than the other groups on the post-test (there were no significant differences between the other three groups). Furthermore, questionnaires given to the students at the end of the sessions showed that the students given grades and praise scored far higher than the 'comments' or the 'no feedback' groups on measures of ego-involvement while those given comments scored higher than the other three groups on measures of task-involvement. Interestingly, those given praise had the highest perceptions of success, even though they had been significantly less successful than the 'comments' group. This is consistent with the findings of Cameron & Pierce (1994), who found that while verbal praise and supportive feedback can increase students' interest in and attitude towards a task, such feedback has little, if any, effect on performance.

These ideas are similar to the framework proposed by Deci & Ryan (1994), who identify four kinds of regulation of behaviour: external, introjected, identified and integrated. External regulation 'describes behaviours that are regulated by contingencies overtly external to the individual', (p. 6), while introjected regulation 'refers to behaviours that are motivated by internal prods and pressures such as self-esteem-relevant contingencies' (p. 6). Identified regulation 'results when a behaviour or regulation is adopted by the self as personally important or valuable' (p. 6), although the motivation is extrinsic, while integrated regulation 'results from the integration of identified values and regulations into one's coherent sense of self' (p. 6). These four kinds of regulation can therefore be regarded as the result of crossing the locus of the value system with that of the motivation, as shown in Fig. 1.

Within this framework, it can be seen that both internal and external motivation can be effective, but only when associated with internally, as opposed to externally, valued aims. Strategies for promoting intrinsic motivation are discussed by Lepper & Hodell (1989).

Related to these findings is the large body of work on the way that students attribute reasons for success and failure, and in particular the work of Dweck and her associates (see Dweck, 1986 for a summary). The crucial variables appear to be:

personalisation	(whether the factors are internal or external);
permanence	(whether the factors are stable or unstable);
specificity	(whether the factors are specific and isolated or whether they are global, generalisable and transferable).

The clear message from the research on attribution theory (see for example Vispoel & Austin, 1995) is that teachers must aim to inculcate in their students the idea that success is due to internal, unstable, specific factors such as effort, rather than on stable general factors such as ability (internal) or whether one is positively regarded by the teacher (external).

Task Motivation Processes

In contrast to those interventions that cue attention to meta-task processes, feedback interventions that direct attention towards the task itself are generally much more successful. Bangert-Drowns *et al.* (1991a) used meta-analysis to condense the findings of 40 studies into the effects of feedback in what they called 'test-like' events (e.g. evaluation questions in programmed learning materials, review tests at the end of a block of teaching, etc.). This study has already been discussed in the section on the quality of feedback. As pointed out there, it was found that providing feedback in the form of answers to the review questions was effective only when students could not 'look ahead' to the answers before they had attempted the questions themselves what Bangert-Drowns *et al.* (1991a), called 'controlling for pre-search availability'). Furthermore, feedback was more effective when the feedback gave details of the correct answer, rather than simply indicating whether the student's answer was correct or incorrect (see also Elshout-Mohr, 1994). Controlling for these two factors eliminated almost all of the negative effect sizes that Bangert-Drowns *et al.* (1991a) found, yielding a mean effect size across 30 studies of 0.58. They also found that the use of pre-tests lowered effect sizes, possibly by giving learners practice in, or by acting as primitive advance organisers for, the material to be covered. They concluded that the key feature in effective use of feedback is that it must encourage 'mindfulness' in the student's response to the feedback. Similar reviews by Dempster (1991, 1992) confirm these findings, but also show that it is important for the interval between successive tests to increase, with the first test occurring shortly after the relevant instruction, but that the effectiveness of successive tests is reduced if the students do not feel successful on the first test. Another important finding in Dempster's work is that tests promote learning as well as sampling it, thus contradicting the often quoted analogy that 'weighing the pig does not fatten it'.

Also discussed in the section on the Quality of feedback was Elawar & Corno's (1985) study of 18 primary school teachers, where it was found that the differences due to being given specific comments on errors and suggestions for strategies, compared with being given just marks, were as great as the differences in achievement due to prior attainment—a significant finding given the well-attested role of previous attainment in determining future success.

Task Learning Processes

What is surprising from reviewing the literature is how little attention has been paid to task characteristics in looking at the effectiveness of feedback. The quality of the feedback intervention, and in particular, how it relates to the task in hand, is crucial.

Feedback appears to be less successful in 'heavily-cued' situations such as are found

in computer-based instruction and programmed learning sequences, and relatively more successful in situations requiring 'higher-order' thinking such as unstructured tests and comprehension exercises (Bangert-Drowns *et al.*, 1991b) or concept mapping (Bernard & Naidu, 1992). Why this might be so is not clear, but one clue comes from a study carried out by Simmons & Cope (1993). In this study, pairs of children, aged 9–11, with little or no experience of Logo programming, showed higher levels of response (as measured by the SOLO taxonomy) when working on angle and rotation problems on paper than when working in a Logo environment, which the authors attributed to the propensity of the immediate feedback given in the Logo environment to encourage incremental or 'trial and improvement' strategies.

Day & Cordon's (1993) study of two 3rd grade classes found that students given a 'scaffolded' response—given as much or as little help as they needed—out-performed those students given a complete solution as soon as they got stuck, and were more able to apply their knowledge to similar, or only slightly related, tasks. Similar results were reported by Declos & Harrington (1991) for students who had used booklets of adjunct questions in order to monitor their progress in tackling practice problems. Improving students' skills in asking for and giving help also has direct positive effects on achievement (Bland & Harris, 1990; Ross, 1995).

However, the *kind* of help is important too. Some researchers have found that repeated explanation of techniques that have previously led to failure is less effective than using alternative strategies (Fuchs *et al.*, 1991), although Mory (1992) suggests that the results are inconclusive on this point. There is also evidence that the quality of dialogue in a feedback intervention is important (Graesser *et al.*, 1995) and can, in fact, be more significant than prior ability and personality factors combined (Clarke, 1988).

Furthermore, while focusing on process goals leads to greater achievement gains than a focus on product goals, feedback related to progress seems to be more effective than feedback on absolute levels of performance (Schunk & Rice, 1991; Schunk & Swartz, 1993a).

In all this, it is easy to gain the impression that formative assessment is a static process of measuring the amount of knowledge currently possessed by the individual, and feeding this back to the individual in some way. However, as the meta-analysis of Fuchs & Fuchs (1986) showed, the effectiveness depends strongly on the systematic analysis and use of feedback by teachers. Furthermore, the account by Lidz (1995) of the history and literature of dynamic assessment (and particularly the work of Vygotsky and Feuerstein) makes plain that formative assessment is as much concerned with prediction (i.e. what someone can learn) as with what they have already learnt, and it is only in interaction with the learner (and the learning) that useful assessments can be made.

Prospects for the Theory and Practice of Formative Assessment

No Meta-analysis

It might be seen desirable, and indeed might be anticipated as conventional, for a review of this type to attempt a meta-analysis of the quantitative studies that have been reported. The fact that this hardly seems possible prompts a reflection on this field of

research. Several studies which are based on meta-analyses have provided useful material for this review. However, these have been focussed on rather narrow aspects of formative work, for example the frequency of questioning. The value of their generalisations is also in question because key aspects of the various studies that they synthesise, for example the quality of the questions being provided at the different frequencies, is ignored because most of the researchers provide no evidence about these aspects.

Individual quantitative studies which look at formative assessment as a whole do exist, and some have been discussed above, although the number with adequate and comparable quantitative rigour would be of the order of 20 at most. However, whilst these are rigorous within their own frameworks and purposes, and whilst they show some coherence and reinforcement in relation to the learning gains associated with classroom assessment initiatives, the underlying differences between the studies are such that any amalgamations of their results would have little meaning.

At one level, these differences are obvious on casual inspection, because each study is associated with a particular pedagogy, with its attendant assumptions about learning: one that in many cases has been constructed as the main element of the innovation under study. There are however deeper differences: even where the research studies appear to be similar in the procedures involved, they differ in the nature of the data which may have been collected—or ignored. The fact that important determining features are often given no attention is one sign of the inadequate conceptualisation of the issues involved, indicating a need for further theory building. From the evidence presented in this review, it is clear that a great deal of theory building still needs to take place in the area of formative assessment, and we shall make suggestions below about a basis for this development.

An underlying problem, which we have already noted in an earlier paper (Wiliam & Black, 1996), is that the term 'formative assessment' is not common in the assessment literature. Such meaning as we have attached to the term here is also represented for others by such terms as 'classroom evaluation', 'curriculum-based assessment', 'feedback', 'formative evaluation' and so on.

Taking further the argument in the section on feedback, we propose, for the sake of simplicity, that the term feedback be used in its least restrictive sense, to refer to any information that is provided to the performer of any action about that performance. This need not necessarily be from an external source (as, for example, would be required by Kluger & DeNisi, 1996), nor need there necessarily be some reference standard against which the performance is measured, let alone some method of comparing the two. The actual performance can be evaluated either in its own terms, or by comparing it with a reference standard. The comparison can either be in terms of equality (i.e. these are the same or different?), as a distance (how far short of—or indeed beyond—the standard was it?) or as diagnosis (what do I need to do to get there?). Adopting the definition (although not the term) proposed by Sadler (1989), we would argue that the feedback in any assessment serves a *formative* function only in the latter case. In other words, assessment is formative only when comparison of actual and reference levels yields information which is then used to alter the gap. As Sadler remarks, 'If the information is simply recorded, passed to a third party who lacks

either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed' (Sadler, 1989, p. 121). In such a case, while the assessment might be formative in purpose, it would not be formative in function and in our view this suggests a basis for distinguishing formative and summative functions of assessment.

Gipps (1994, chapter 9) draws attention to a paradigm shift from a testing culture to an assessment culture, associated with a shift from psychometrics to the assessment of learning. Similarly, Shinn & Good III (1993) argue that there needs to be a 'paradigm shift' in assessment, from what they call the current assessment paradigm (and what we have here called summative functions of assessment) to what they call the 'problem-solving paradigm' (broadly equivalent to what we are here calling the formative functions of assessment). They illustrate the distinction by the differences in the way that questions are posed in the two paradigms along various dimensions (see Table 1—from Shinn & Hubbard, 1992). Summative functions of assessment are concerned with consistency of decisions across (relatively) large groups of students, so that the over-riding imperative is that meanings are shared by different users of assessment results. A particular problem for the constructors of summative assessments is that exactly who will be making use of the assessment results is likely to be undetermined. In contrast, formative functions of assessment prioritise desirable consequences either for (relatively) small groups of students (such as a teaching group) or for particular individuals.

The lack of clarity about the formative/summative distinction is more or less evident in much of the literature. Examples can be found in the flourish of articles and books, notably in the USA, about performance assessment, authentic assessment, portfolio assessment and so on, where innovations are described, sometimes with evidence which is presented as an evaluation, with the focus only on the reliability of the teachers' assessments and the feasibility of the classroom work involved. What is often missing is a clear indication as to whether the innovation is meant to serve the short-term purpose of improvement of learning, or the long-term purpose of providing a more valid form of summative assessment, or both.

The Theoretical Basis

All that can be set out here are a few 'notes towards a theory of formative assessment', which are offered partly because they may be a helpful aid to reflection on the work surveyed and partly because they may be helpful in looking ahead to the implications of this work.

Two key contributions, to which reference has already been made, are those of Sadler (1989) and Tittle (1994). Sadler built upon Ramaprasad's notion of the gap between the state revealed by feedback and the desired state, emphasising that action will be inhibited if this gap is seen as impracticably wide. He further argued that ultimately, the action to close that gap must be taken by the student—a student who automatically follows the diagnostic prescription of a teacher without understanding of its purpose or orientation will not learn. Thus self-assessment by the student is not an interesting

option or luxury; it has to be seen as essential. Given this, the orientation by a student of his or her work can only be productive if that student comes to share the teacher's vision of the subject matter. Some (e.g. Klenowski, 1995) argue that this can be done by clarifying objectives, but others (e.g. Claxton, 1995; Wiliam, 1994) argue that these definitions must remain implicit if they are not to distort learning.

A development of this theory seems to call for links to compatible learning theories and to theories of the meta-cognition and locus of control of the learner.

Tittle's (1994) framework emphasises three dimensions. The first, the epistemology and theories involved, can relate both to positions held in relation to learning in general, and to the particular epistemology relevant to the subject matter concerned. The nature of the epistemology, and so of the meta-cognition involved, in (say) aesthetic appreciation of poetry will be very different from that for (say) physics, and hence many features of formative assessment will differ between these two fields of learning. The second dimension is the more evident one of the assessment characteristics; it can be remarked here that in several of the studies reported here, little is said about the detail of these, or about the distinctive effects of the particular subject matter involved.

Tittle's third dimension brings in the interpreter and user, and she particularly stresses the importance of these. In relation to students, this emphasis is reinforced and developed by Sadler's arguments, but the teacher's beliefs, about the subject matter, about learning, and about the students and the class, must also be important components in any model, if only because it is on the basis of these that appraisals of Sadler's 'gap' must be formulated. Tittle also makes the important point that while modern conceptions of validity theory (e.g. Messick, 1989) stress the value-laden nature of assessment processes, the actual nature of those values is excluded, creating the impression that one (presumably reasonably coherent) set of values is as good as any other. Thus current conceptions of validity provide no guide as to what 'ought' to be going on, merely a theoretical framework for discussing what *is* going on.

This emphasis on the ethical and moral aspects of assessment is a feature of the perspective outlined by Aikenhead (1997). He draws upon the work of Habermas (1971, p. 308) and Ryan (1988) to propose that consideration of assessment can fall within three paradigms that are commonly encountered in the social sciences. One, the empirical-analytic, clearly links to the psychometric emphasis in standardised testing. The second, the interpretative paradigm, has to be adopted in formative assessment, and this link brings out the importance of understanding a learner's response in relation to that learner's expectations and assumptions about the classroom process, together with his or her interpretation of the task demand and of the criteria for success. In the third, the critical-theoretic paradigm, one would seek a critique of the wider purposes being pursued, notably the empowerment of the learner, and the choice between either selecting an elite or achieving excellence for all. This paradigm also calls into play the need for a critique of the learning goals (and of the assessment criteria through which they are operationalised) which should ask whose interests these goals are designed to serve.

Similar concerns motivate the theoretical framework proposed by Davis (1997) as a result of a detailed study of the changes (over a two-year period) of the practice of a single middle-school mathematics teacher in the way she reacted to students'

responses to her questions. Initially, the teacher's reactions tended to focus on the extent to which the student responses accorded with the teacher's expectations (what Davis terms 'evaluative' listening). After sustained reflection and discussion with the researcher over a period of several months, the teacher's reaction placed increasing emphasis on 'information-seeking' as opposed to the 'response-seeking' which characterised the earlier lessons ('interpretive' listening). Towards the end of the two-year period, there was a further shift in the teacher's practice, with a marked move away from clear lesson structures and pre-specified learning outcomes, and towards the exploration of potentially rich mathematical situations, in which the teacher is a co-participant. Most notably, in this third phase, the teacher's own views of the subject matter being 'taught' developed and altered along with that of the students ('hermeneutic' listening). It is clear therefore that a commitment to the use of formative assessment necessarily entails a move away from unitary notions of intelligence (Wolf *et al.*, 1991).

Expectations and the Social Setting

These last two analyses bring out a feature which in our view has been absent from a great deal of the research we have reviewed. This is that all the assessment processes are, at heart, social processes, taking place in social settings, conducted by, on and for social actors. Guy Brousseau (1984) has used the term 'didactical contract' to describe the network of (largely implicit) expectations and agreements that are evolved between students and teachers. A particular feature of such contracts is that they serve to delimit 'legitimate' activity by the teacher. For example, in a classroom where the teacher's questioning has always been restricted to 'lower-order' skills, such as the production of correct procedures, students may well see questions about 'understanding' or 'application' as unfair, illegitimate or even meaningless (Schoenfeld, 1985).

As Tittle's (1994) approach emphasises, the 'opening moves' of teachers and students in the negotiation of such classroom contracts will be determined by their epistemological, psychological and pedagogical beliefs. For example, when a teacher questions a student, the teacher's beliefs will influence both the questions asked and the way that answers are interpreted. An important principle here is the distinction between 'fit' and 'match' (von Glasersfeld, 1987, p. 13). For example, a teacher may set student problems in solving systems of simple equations. If students answer all the questions correctly, the teacher may well conclude that the students have 'understood' the topic, i.e. they assume that the students' understanding matches theirs. However, this is frequently not the case. For example, when asked to solve the following two equations

$$3a = 24$$
$$a + b = 16$$

many students believe that it is impossible, saying things like 'I keep getting b is 8, but it can't be because a is 8'. This is because in the examples encountered in most textbooks, each letter stands for a different number. The students' understanding is therefore not a match but only a 'fit' with the teacher's. The relationship between fit and match depends critically on the richness of the questions used by the teacher, and

this, in turn will depend on the teacher's subject knowledge, their theories of learning, and their experience of learners.

A study of seven experienced elementary school teachers examined the implicit criteria that teachers used to determine whether students had 'understood' something (Reynolds *et al.*, 1995). After studying and discussing video extracts and transcripts of lessons, seven 'indicators of understanding' emerged which were agreed by all seven teachers, although they were regarded not as a static check-list, but rather as a series of potential clues to the level of the student's understanding:

(1) changes in demeanour: students who had understood were 'bright-eyed' while those who had not appeared half-hearted;
(2) extension of a concept: students who have understood something often take the idea further on their own initiative;
(3) making modifications to a pattern: students who understand, spontaneously start making their own modifications, while those who don't understand imitate or follow rules;
(4) using processes in a different context: students who have understood a particular idea often start seeing the same patterns elsewhere;
(5) using shortcuts: only students who are sure of the 'big picture' can short-cut a procedure so that thinking up or using a short-cut is taken as evidence of understanding;
(6) ability to explain: students who have understood something are usually able to explain it;
(7) ability to focus attention: persistence on a task is taken as a sign of understanding.

It may be that some teachers are content with 'fits' rather than 'matches' because they are unaware of the possibilities for students' conceptions that are different from their own. However, it seems likely that most teachers are aware of the benefits of richer questioning styles, but find that such approaches are difficult to implement in 'real classrooms' (Dassa, 1990). In this respect, computer software that enables teachers to provide formative and diagnostic feedback may have a role to play (Dassa *et al.*, 1993; Wiliam, 1997), although there is little evidence so far about the actual benefits of such software.

In turn, the student's responses to questioning will depend on a host of factors. Whether the student believes ability to be incremental or fixed will have a strong influence on how the student sees a question—as an opportunity to learn or as a threat to self-esteem (Dweck, 1986). Even where the student has a 'learning' as opposed to 'performance' orientation, the student's belief about what counts as 'academic work' (Doyle, 1988) will have a profound impact on the 'mindfulness' with which that student responds. The study of two middle-school teachers by Lorsbach *et al.* (1992) cited in the earlier section on Current practice found that a major threat to the validity of test-result interpretation was the extent to which students could construct meanings for the tasks they were set, and the extent to which teachers could construct meanings for the students' responses. They also found that teachers used assessment results as if they gave information on what students knew, whereas, in fact, they were better indicators of motivation and task completion.

More specifically, the actual context of the assessment can also influence what students believe is required. An example is a study of a grade 5 geometry class (Hall *et al.*, 1995) where performance was assessed in two ways—via a multiple choice test and with an assignment in which students had to design a HyperCard geometry tutorial. In the multiple choice test, the students focused on the grades awarded, while in the tutorial task, students engaged in much more presentation and qualitative discussion of their work. Perhaps most significantly, discussion amongst students of the different tutorials focused much more directly on the subject matter (i.e. geometry) than did the (intense) comparison of grades on the multiple choice test.

The actions of teachers and students are also 'enframed' (Mitchell, 1991) by the structures of schools and society and typically knowledge is closely tied to the situation in which it is learnt (Boaler, 1997). Spaces in schools are designated for specified activities, and given the importance attached to 'orderliness' in most classrooms, teachers' actions are as often concerned with establishing routines, order and student satisfaction as they are with developing the student's capabilities (Torrance & Pryor, 1995; Pryor & Torrance, 1996). A review by Rismark (1996) shows that students are frequently marginalised and their work undervalued if they use frames of reference from their personal experiences outside school and Filer (1993) found that children learning handwriting and spelling in English primary school classrooms were constrained by the teacher to develop these skills in standard contexts, so that their own personal experiences were 'blocked out'. In this way, formal, purportedly 'objective' assessments made by teachers may be little more than the result of successive sedimentation of previous 'informal' assessments—in extreme cases the self-fulfilling prophecy of teachers' labelling of students (Filer, 1995).

In trying to reconcile these effects of structure and agency, Bourdieu's notion of habitus (Bourdieu, 1985) may be particularly fruitful. Traditional approaches to sociological analysis have used coarse categories such as gender, race, and social class to 'explain' differences in, for example, outcomes, thus tending to treat all those within a category as being homogenous. Bourdieu uses the notion of habitus to describe the orientations, experiences and positions adopted by social actors, particularly in order to account for the differences between individuals in the same categories. Such a notion seems particularly appropriate for describing classrooms, in view of the fact that the experiences of students in the same classroom can be so different (Dart & Clarke, 1989).

Research—prospects and needs

The above discussion has clear implications for the design of research investigations. It draws attention to the range of important features which will combine to determine the effects of any classroom regime. In the light of such a specification, it is clear that most of the studies in the literature have not attended to some of the important aspects of the situations being researched. A full list of important and relevant aspects would include the following:

- the assumptions about learning underlying the curriculum and pedagogy;
- the rationale underlying the composition and presentation of the learning work;

- the precise nature of the various types of assessment evidence revealed by the learner's responses;
- the interpretative framework used by both teachers and learners in responding to this evidence;
- the learning work used in acting on the interpretations so derived;
- the divisions of responsibility between learners and teachers in these processes;
- the perceptions and beliefs held by the learners about themselves as learners about their own learning work, and about the aims and methods for their studies;
- the perceptions and beliefs of teachers about learning, about the 'abilities' and prospects of their students, and about their roles as assessors;
- the nature of the social setting in the classroom, as created by the learning and teaching members and by the constraints of the wider school system as they perceive and evaluate them;
- issues relating to race, class and gender, which appear to have received little attention in research studies of formative assessment;
- the extent to which the context of any study is artificial and the possible effects of this feature on the generalisability of the results.

To make adequate report of all of these, let alone control them in any classical quantitative design, would seem very difficult. This is not to imply that reliable measures of outcomes, both of learning and of attitudes to the subjects learnt, are not to be sought—although one of the problems evident in many of the studies seems to be that although they are serving learning aims that the established methods ignore or play down, they have to justify themselves in relation to tests which are adapted to the established methods only. There is clearly a need for a combination of such measures with richer qualitative studies of processes and interactions within the classroom. If, as we believe, there is a need to evolve new approaches as quickly as possible, such studies might well focus on the problems of change and attendant disorientations.

Particular attention ought to be paid to two specific problems. The first is the evidence in many studies that new emphasis on formative assessment is of particular benefit to the disadvantaged and low-attaining learners—evidence which is not supported in the results of other studies. The apparent contradictions here probably arise because there are some important features of the classrooms that have yet to be recorded and understood. If it is true that the ranges of school achievement might be narrowed by the enhancement of the achievement of those hitherto seen as slow learners, then there are very strong social and educational reasons for giving high priority to sensitive research and development work to see how to understand and tackle the issues involved.

The second problem, or clutch of problems, relates to the possible confusions and tensions, both for teachers and learners, between the formative and summative purposes which their work might have to serve. It is inevitable that all will be involved, one way or the other, in working to both purposes, and if an optimum balance is not sought, formative work will always be insecure because of the threat of renewed dominance by the summative.

TABLE I. Different questions arising from Paradigm Shift (Shinn & Hubbard, 1992)

Dimension	Current assessment paradigm	Problem-solving paradigm
Purpose	Do assessment results spread out individuals facilitating *classification/ placement* into groups?	Does assessment result in socially meaningful *student outcomes* for the individual?
Test Validity	Does the assessment device measure what it says it measures? Criterion-related Validity: Does the test correlate with other tests purporting to increase the same thing? Construct Validity: Does the test display a stable factor structure?	Are the *inferences* and *actions* based on test scores *adequate* and *appropriate* (Messick, 1989)? Treatment Validity: Do decisions regarding target behaviors and treatments based on knowledge obtained from the assessment procedure result in *better student outcomes* than decisions based on alternative procedures (Hayes *et al.*, 1983)?
Unit of Analysis	Groups: Probabilistic statements about individuals: Do students with similar assessment results *most likely* display similar characteristics?	Individuals: Does assessment show that *this* treatment is working for *this* student?
Time Line	Summative: Does the assessment indicate whether or not the intervention *did* work?	Formative: Does the assessment indicate that *this* treatment is working for *this* student?
Level of Inference	Does the assessment provide an *indirect* measure of an unobservable construct?	Does the assessment *directly* measure important target behaviors or skills?
Locus of the Problem	Does the assessment identify relevant *student characteristics* that contribute to problem etiology?	Does assessment identify relevant *curriculum, instruction and contextual* factors [that] contribute to problem solution?
Focus	Problem Certification: Does assessment accurately identify *problems*?	Problem Solution: Does the assessment accurately identify *solutions*?
Test Reliability	Are test scores stable over time? Are scores based on different behavior samples, obtained in different contexts/settings consistent?	What factors account for the variability in student performance?
Context	Does the assessment provide a comparison with students receiving a nationally representative range of curriculum and instruction?	Does the assessment provide a comparison with students receiving comparable curriculum and instruction?
Dimension of dependent variable	Does the assessment provide information regarding the *level* of pupil performance?	Does the assessment provide information regarding the *level* of pupil performance and the *slope* of pupil progress?

Are there Implications for Policy?

Table I could be read alongside the section on Strategies and tactics for teachers as helping to determine the essential elements of any strategy to improve learning through thorough implementation of formative assessment. These elements would be the setting of clear goals, the choice, framing and articulation of appropriate learning tasks, the deployment of these with appropriate pedagogy to evoke feedback, noting the arguments in the section on Students and formative assessment, and the appropriate interpretation and use of that feedback to guide the learning trajectory of students. Within and running through any such plan should be a commitment to involving students in the processes of self- and peer-assessment as emphasised in the above section, underpinned by a constructivist approach to learning.

There are clearly many different ways in which such guidelines could be incorporated into classroom practice, and whilst the various experiments and schemes described throughout this review, and the particular strategies explored in the section on Systems, give useful examples, there is clearly no single royal road. In particular, a careful reading of the earlier sections on Reception and response, on Goal orientation and Self-perception on students' response to feedback, and of parts of sections on The quality of feedback, Feedback and Expectations and the social setting, should show that in framing the feedback that they give to students teachers have to keep in mind several important and delicate considerations which are neither widely known nor understood.

For public policy towards schools, the case to be made here is firstly that significant learning gains lie within our grasp. The research reported here shows conclusively that formative assessment does improve learning. The gains in achievement appear to be quite considerable, and as noted earlier, amongst the largest ever reported for educational interventions. As an illustration of just how big these gains are, an effect size of 0.7, if it could be achieved on a nationwide scale, would be equivalent to raising the mathematics attainment score of an 'average' country like England, New Zealand or the United States into the 'top five' after the Pacific rim countries of Singapore, Korea, Japan and Hong Kong (Beaton *et al.*, 1996).

If this first point is accepted, then the second move is for teachers in schools to be provoked and supported in trying to establish new practices in formative assessment, there being extensive evidence to show that the present levels of practice in this aspect of teaching are low (Black, 1993b; McCallum *et al.*, 1993), and that the level of resources devoted to its support, at least in the UK since 1988, has been almost negligible (Daugherty, 1995).

There is no doubt that, whilst building coherent theories, adequate descriptions, and firmly grounded guides to practice, for formative assessment is a formidable undertaking, there is enough evidence in place for giving helpful guidance to practical action (for an account of a major state-wide assessment system which incorporates formative and summative functions of assessment, see, for example, Rowe & Hill, 1996). Furthermore, despite the existence of some marginal and even negative results, the range of conditions and contexts under which studies have shown that gains can be achieved must indicate that the principles that underlie achievement of substantial improvements in learning are robust. Significant gains can be achieved by many

different routes, and initiatives here are not likely to fail through neglect of delicate and subtle features.

This last point is very important because there does not emerge, from this present review, any one optimum model on which such a policy might be based. What does emerge is a set of guiding principles, with the general caveat that the changes in classroom practice that are needed are central rather than marginal, and have to be incorporated by each teacher into his or her practice in his or her own way (Broadfoot *et al.*, 1996). That is to say, reform in this dimension will inevitably take a long time, and need continuing support from both practitioners and researchers.

Acknowledgements

The compilation of this review was made possible by the provision of a grant from the Nuffield Foundation, whose support we gratefully acknowledge. We would also like to thank the members of the British Educational Research Association's Assessment Policy Task Group who commissioned the work, provided helpful comments on earlier drafts, and made suggestions for additional references for inclusion. Despite such support, this review is bound to contain errors, omissions and misrepresentations, which are, of course, the entire responsibility of the authors.

References

ADELMAN, C., KING, D. & TREACHER, V. (1990) Assessment and teacher autonomy, *Cambridge Journal of Education*, 20, pp. 123–133.

AIKENHEAD, G. (1997) A framework for reflecting on assessment and evaluation, in: *Globalization of Science Education—papers for the Seoul International Conference*, pp. 195–199 (Seoul, Korea, Korean Educational Development Institute).

ALLINDER, R.M. (1995) An examination of the relationship between teacher efficacy and curriculum-based measurement and student-achievement, *Remedial and Special Education*, 16, pp. 247–254.

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

AMES, C. & ARCHER, J. (1988) Achievement goals in the classroom: students' learning strategies and motivation process, *Journal of Educational Psychology*, 80, pp. 260–267.

ARTHUR, H. (1995) Student self-evaluations—how useful—how valid, *International Journal of Nursing Studies*, 32, pp. 271–276.

BACHOR, D.G. & ANDERSON, G.O. (1994) Elementary teachers' assessment practices as observed in the Province of British Columbia, Canada, *Assessment in Education*, 1, pp. 63–93.

BAIRD, J.R., FENSHAM, P.J., GUNSTONE, R.F. & WHITE, R T. (1991) The importance of reflection in improving science teaching and learning, *Journal of Research in Science Teaching*, 28, pp. 163–182.

BANGERT-DROWNS, R.L., KULIK C.-L.C., KULIK, J.A. & MORGAN, M.T. (1991a) The instructional effect of feedback in test-like events), *Review of Educational Research*, 61, pp. 213–238.

BANGERT-DROWNS, R.L., KULIK, J.A. & KULIK, C.-L.C. (1991b) Effects of frequent classroom testing, *Journal of Educational Research*, 85, pp. 89–99.

BEATON, A.E., MULLIS, I.V.S., MARTIN, M.O., GONZALEZ, E.J., KELLY, D.L. & SMITH, T.A. (1996) *Mathematics Achievement in the Middle School Years* (Boston, MA, Boston College).

BELANOFF, P. & DICKSON, M. (Eds) (1991) *Portfolios: process and product* (Portsmouth, NH, Boynton/Cook).

BENNETT, S.N., WRAGG, E.C., CARRE, C.G. & CARTER, D.G.S. (1992) A longitudinal study of primary teachers perceived competence in, and concerns about, national curriculum implementation, *Research Papers in Education*, 7, pp. 53–78.

BENOIT, J. & YANG, H. (1996) A redefinition of portfolio assessment based on purpose: findings and implications from a large scale program, *Journal of Research and Development in Education*, 29, pp. 181–191.

BERGAN, J.R., SLADECZEK, I.E., SCHWARZ, R.D. & SMITH, A.N. (1991) Effects of a measurement and planning system on kindergartners' cognitive development and educational programming, *American Educational Research Journal*, 28, pp. 683–714.

BERNARD, R.M. & NAIDU, S. (1992) Post-questioning, concept mapping and feedback—a distance education field experiment, *British Journal of Educational Technology*, 23, pp. 48–60.

BLACK, H.D. & DOCKRELL, W.B. (1984) *Criterion-referenced assessment in the classroom* (Edinburgh, Scottish Council for Research in Education).

BLACK, P. & ATKIN, J.M. (1996) *Changing the subject: innovations in science, mathematics and technology education* (London, Routledge with OECD).

BLACK, P.J. (1993a) Assessment policy and public confidence: comments on the BERA Policy Task Group's article 'Assessment and the improvement of education', *The Curriculum Journal*, 4, pp. 421–427.

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

BLAND, M. & HARRIS, G. (1990) Peer tutoring, *School Science Review*, 71(255), pp. 142–144.

BLOCK, J.H. & BURNS, R.B. (1976) Mastery learning, in: L.S. SHULMAN (Ed.) *Review of Research in Education* (Itasca, IL, Pocock).

BLUMENFELD, P.C. (1992) Classroom learning and motivation: clarifying and expanding goal theory, *Journal of Educational Psychology*, 84, pp. 272–281.

BOALER, J. (1997) *Experiencing school mathematics: teaching styles, sex and setting* (Buckingham, UK Open University Press).

BOL, L. & STRAGE, A. (1996) The contradiction between teachers' instructional goals and their assessment practices in high school biology courses, *Science Education*, 80, pp. 145–163.

BONNIOL, J.J. (1991) The mechanisms regulating the learning process of pupils: contribution to a theory of formative assessment, in: P. WESTON (Ed.) *Assessment of Pupils' Achievement: Motivation and School Success*, pp. 119–137 (Amsterdam, Swets and Zeitlinger).

BOULET, M.M., SIMARD, G. & DEMELO, D. (1990) Formative evaluation effects on learning music, *Journal of Educational Research*, 84, pp. 119–125.

BOURDIEU, P. (1985) The genesis of the concepts of 'habitus' and 'field', *Sociocriticism*, 2(2), pp. 11–24.

BROADFOOT, P. (1992) Multilateral evaluation: a case study of the national evaluation of records of achievement (PRAISE) project, *British Educational Research Journal*, 18, pp. 245–260.

BROADFOOT, P., JAMES, M., MCMEEKING, S., NUTTALL, D. & STIERER, B. (1990) Records of achievement: report of the National Evaluation of Pilot Schemes, in: T. HORTON (Ed.) *Assessment Debates*, pp. 87–103 (London, Hodder and Stoughton).

BROADFOOT, P., OSBORN, M., PANEL, C. & POLLARD, A. (1996) Assessment in French primary schools, *Curriculum Journal*, 7, pp. 227–246.

BROMME, R. & STEINBERG, H. (1994) Interactive development of subject matter in mathematics classrooms, *Educational Studies in Mathematics*, 27, pp. 217–248.

BROUSSEAU, G. (1984) The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics, in: H.-G. STEINER (Ed.) *Theory of Mathematics Education: ICME 5 topic area and miniconference*, pp. 110–119 (Bielefeld, Germany, Institut für Didaktik der Mathematik der Universität Bielefeld).

BROWN, A.L., CAMPIONE, J.C., WEBBER, L.S. & MCGILLY, K. (1992) Interactive learning environments: a new look at assessment and instruction, in: B.R. GIFFORD & M.C. O'CONNOR (Eds) *Changing Assessments: alternative views of aptitude, achievement and instruction*, pp. 121–211 (Boston USA and Dordrecht Netherlands, Kluwer).

BROWN, M. & DENVIR, B. (1987) The feasibility of class administered diagnostic assessment in primary mathematics, *Educational Research*, 29, pp. 95–107.

BROWN, R., PRESSLEY, M., VAN METER, P. & SCHUDER, T. (1996) A quasi-experimental validation of transactional strategies instruction (with low-achieving second-grade readers, *Journal of Educational Psychology*, 88, pp. 18–37.

BUTLER, D.L. & WINNE, P.H. (1995) Feedback and self-regulated learning: a theoretical synthesis, *Review of Educational Research*, 65, pp. 245–281.

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

BUTLER, J. & BEASLEY, W. (1987) The impact of assessment changes on the science curriculum, *Research in Science Education*, 17, pp. 236–243.

BUTLER, R. (1987) Task-involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest and performance, *Journal of Educational Psychology*, 79, pp. 474–482.

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.

BUTLER, R. & NEUMAN, O. (1995) Effects of task and ego-achievement goals on help-seeking behaviours and attitudes, *Journal of Educational Psychology*, 87, pp. 261–271.

CALFEE, R.C. & FREEDMAN, S.W. (1996) Classroom writing portfolios: old, new, borrowed, blue, in: R. CALFEE & P. PERFUMO (Eds) *Writing Portfolios in the Classroom*, pp. 3–26 (Mahwah, NJ, Lawrence Erlbaum).

CALFEE, R.C. & PERFUMO, P. (Eds) (1996a) *Writing Portfolios in the Classroom* (Mahwah, NJ, Lawrence Erlbaum).

CALFEE, R. C. & PERFUMO, P. (1996b) A national survey of portfolio practice: what we learned and what it means, in: R. CALFEE & P. PERFUMO (Eds) *Writing Portfolios in the Classroom*, pp. 63–81 (Mahwah, NJ, Lawrence Erlbaum).

CAMERON, J. & PIERCE, D.P. (1994) Reinforcement, reward, and intrinsic motivation: a meta-analysis, *Review of Educational Research*, 64, pp. 363–423.

CARLSEN, W.S. (1991) Questioning in classrooms—a sociolinguistic perspective, *Review of Educational Research*, 61, pp. 157–178.

CARROLL, W. M. (1994) Using worked examples as an instructional support in the algebra classroom, *Journal of Educational Psychology*, 83, pp. 360–367.

CAVENDISH, S., GALTON, M., HARGREAVES, L. & HARLEN, W. (1990) *Observing Activities* (London, Paul Chapman).

CIZEK, G.J., FITZGERALD, S.M. & RACHOR, R.E. (1995) Teachers' assessment practices: preparation, isolation and the kitchen sink, *Educational Assessment*, 3, pp. 159–179.

CLARKE, J. (1988) Classroom dialogue and science achievement, *Research in Science Education*, 18, pp. 83–94.

CLAXTON, G. (1995) What kind of learning does self-assessment drive? Developing a 'nose' for quality; comments on Klenowski, *Assessment in Education*, 2, pp. 339–343.

COLLINS, A. (1992) Portfolios for science education: issues in purpose, structure and authenticity, *Science Education*, 76, pp. 451–463.

COSGROVE, M. & SCHAVERIEN, L. (1996) Childrens' conversations and learning science and technology, *International Journal of Science Education*, 18, pp. 105–116.

COURTS, P.L. & MCINERNEY, K.H. (1993) *Assessment in Higher Education: politics, pedagogy, and portfolios* (Westport, CT, Praeger/Greenwood).

CRAVEN, R.G., MARSH, H.W. & DEBUS, R.L. (1991) Effects of internally focused feedback on enhancement of academic self-concept, *Journal of Educational Psychology*, 83, pp. 17–27.

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

DARO, P. (1996) Standards and portfolio assessment, in: J.B. BARON & D.P. WOLF (Eds) *Performance-Based Student Assessment: challenges and possibilities*, pp. 239–260 (Chicago, IL, University of Chicago Press).

DART, B.C. & CLARKE, J.A. (1989) Target students in year 8 science classrooms: a comparison with and extension of existing research, *Research in Science Education*, 19, pp. 67–75.

DASSA, C. (1990) From a horizontal to a vertical method of integrating educational diagnosis with classroom assessment, *Alberta Journal of Educational Research*, 36, pp. 35–44.

DASSA, C., VAZQUEZ-ABAD J. & AJAR, D. (1993) Formative assessment in a classroom setting: from practice to computer innovations, *Alberta Journal of Educational Research*, 39, pp. 111–125.

DAUGHERTY, R. (1995) *National Curriculum Assessment. A review of policy 1987–1994* (London, Falmer Press).

DAVIS, B. (1997) Listening for differences: an evolving conception of mathematics teaching, *Journal for Research in Mathematics Education*, 28, pp. 355–376.

DAWS, N. & SINGH, B. (1996) Formative assessment: to what extent is its potential to enhance pupils' science being realized? *School Science Review*, 77(281), pp. 93–100.

DAY, J.D. & CORDON, L.A. (1993) Static and dynamic measures of ability: an experimental comparison, *Journal of Educational Psychology*, 85, pp. 76–82.

DECI, E.L. & RYAN, R.M. (1994) Promoting self-determined education, *Scandinavian Journal of Educational Research*, 38, pp. 3–14.

DELCLOS, V.R. & HARRINGTON, C. (1991) Effects of strategy monitoring and proactive instruction on children's problem-solving performance, *Journal of Educational Psychology*, 83, pp. 35–42.

DEMPSTER, F. N. (1991) Synthesis of research on reviews and tests, *Educational Leadership*, 48(7), pp. 71–76.

DEMPSTER, F.N. (1992) Using tests to promote learning: a neglected classroom resource, *Journal of Research and Development in Education*, 25, pp. 213–217.

DENO, S.L. (1993) Curriculum-based measurement, in: J.J. KRAMER (Ed.) *Curriculum-Based Measurement*, pp. 1–24 (Lincoln, NE, Buros Institute of Mental Measurements).

DOYLE, W. (1988) Work in mathematics classes: the context of students' thinking during instruction, *Educational Psychologist*, 23, pp. 167–180.

DUMAS-CARRE, A. & LARCHER, C. (1987) The stepping stones of learning and evaluation, *International Journal of Science Education*, 9, pp. 93–104.

DUSCHL, R.D. & GITOMER, D.H. (1997) Strategies and challenges to changing the focus of assessment and instruction in science classrooms, *Educational Assessment*, 4, pp. 37–73.

DWECK, C.S. (1986) Motivational processes affecting learning, *American Psychologist (Special Issue: Psychological science and education)*, 41, pp. 1040–1048.

DWIGHT, K. (1988) Building in assessment, *School Science Review*, 70(252), pp. 119–125.

EDWARDS, R. & SUTTON, A. (1991) A practical approach to student-centred learning, *British Journal of Educational Technology*, 23, pp. 4–20.

ELAWAR, M.C. & CORNO, L. (1985) A factorial experiment in teachers' written feedback on student homework: changing teacher behaviour a little rather than a lot, *Journal of Educational Psychology*, 77, pp. 162–173.

ELSHOUT-MOHR, M. (1994) Feedback in self-instruction, *European Education*, 26, pp. 58–73.

FAIRBROTHER, R., BLACK, P.J. & GILL, P. (Eds) (1994) *Teachers Assessing Pupils: lessons from science classrooms* (Hatfield, Association for Science Education).

FERNANDES, M. & FONTANA, D. (1996) Changes in control beliefs in Portuguese primary school pupils as a consequence of the employment of self-assessment strategies, *British Journal of Educational Psychology*, 66, pp. 301–313.

FERRIMAN, B. & LOCK, R. (1989) OCEA The development of a graded-assessment scheme in science, Part IV The pilot phase, *School Science Review*, 70(253), pp. 97–102.

FILER, A. (1993) Contexts of assessment in a primary classroom, *British Educational Research Journal*, 19, pp. 95–107.

FILER, A. (1995) Teacher Assessment: social process and social products, *Assessment in Education*, 2, pp. 23–38.

FONTANA, D. & FERNANDES, M. (1994) Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils, *British Journal of Educational Psychology*, 64, pp. 407–417.

FOOS, P.W., MORA, J.J. & TKACZ, S. (1994) Student study techniques and the generation effect, *Journal of Educational Psychology*, 86, pp. 567–576.

FREDERIKSEN, J.R. & WHITE, B.J. (1997) Reflective assessment of students' research within an inquiry-based middle school science curriculum, paper presented at *the Annual Meeting of the AERA Chicago 1997.*

FUCHS, L.S. (1993) Enhancing instructional programming and student achievement with curriculum-based measurement, in: J.J. KRAMER (Ed.) *Curriculum-Based Measurement*, pp. 65–103 (Lincoln, NE, Buros Institute of Mental Measurements).

FUCHS, L.S. & FUCHS, D. (1986) Effects of systematic formative evaluation: a meta-analysis, *Exceptional Children*, 53, pp. 199–208.

FUCHS, L.S., FUCHS, D., HAMLETT, C.L. & STECKER, P.M. (1991) Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations, *American Educational Research Journal*, 28, pp. 617–641.

FUCHS, L.S., FUCHS, D., HAMLETT, C.L., WALZ, L. & GERMANN, G. (1993) Formative evaluation of academic progress—how much growth can we expect? *School Psychology Review*, 22, pp. 27–48.

GARNETT, P.J. & TOBIN, K. (1989) Teaching for understanding: exemplary practice in high school chemistry, *Journal of Research in Science Teaching*, 26, pp. 1–14.

GEISLER-BRENSTEIN, E. & SCHMECK, R.R. (1995) The revised inventory of learning processes: a multi-faceted perspective on individual differences in learning, in: M. BIRENBAUM & F.J.R.C. DOCHY (Eds) *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*, pp. 283–317 (Boston, Kluwer).

GILBERT, G. (1996) Developing an assessment stance in primary education in England, *Assessment in Education*, 3, pp. 55–74.

GIPPS, C. (1994) *Beyond Testing: towards a theory of educational assessment* (London, Falmer Press).

GIPPS, C., MCCALLUM, B. & BROWN, M. (1997) Models of teacher assessment among primary school teachers in England, *The Curriculum Journal*, 7, pp. 167–183.

GOOD, T.L. & GROUWS, D.A. (1975) *Process product relationships in fourth-grade mathematics classrooms* (Columbia, MO, University of Missouri).

GRAESSER, A.C., PERSON, N.K. & MAGLIANO, J.P. (1995) Collaborative dialogue patterns in naturalistic one-to-one tutoring, *Applied Cognitive Psychology*, 9, pp. 495–522.

GRIFFITHS, M. & DAVIES, G. (1993) Learning to Learn: action research from an equal opportunities perspective in a junior school, *British Educational Research Journal*, 19, pp. 43–58.

GRISAY, A. (1991) Improving assessment in primary schools: 'APER' research reduces failure rates, in: P. WESTON (Ed.) *Assessment of Pupils' Achievement: motivation and school success*, pp. 103–118 (Amsterdam, Swets and Zeitlinger).

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.

GUSKEY, G.R. & GATES, S.L. (1986) Synthesis of research on the effects of mastery learning in elementary and secondary classrooms, *Educational Leadership*, 33(8), pp. 73–80.

GUSKEY, T.R. & PIGOTT, T.D. (1988) Research on group-based mastery learning programs: a meta-analysis, *Journal of Educational Research*, 81, pp. 197–216.

HABERMAS, J. (1971) *Knowledge and Human Interest* (Boston, MA, Beacon).

HALL, R.P., KNUDSEN, J. & GREENO, J.G. (1995) A case study of systemic aspects of assessment technologies, *Educational Assessment*, 3, pp. 315–361.

HALL, K., WEBBER, B., VARLEY, S., YOUNG, V. & DORMAN, P. (1997) A study of teacher assessment at key stage 1, *Cambridge Journal of Education*, 27, pp. 107–122.

HARLEN, W., GIPPS, C., BROADFOOT, P. & NUTTALL, D. (1992) Assessment and the improvement of education, *The Curriculum Journal*, 3, pp. 215–230.

HARLEN, W., MALCOLM, H.C. & BYRNE, M. (1995) Teachers' assessment and national testing in primary schools in Scotland: roles and relationships, *Assessment in Education*, 2, pp. 126–144.

HARLEN, W. & MALCOLM, H. (1996) Assessment and testing in Scottish primary schools, *The Curriculum Journal*, 7, pp. 247–257.

HARNETT, P. (1993) Identifying progression in children's understanding: the use of visual materials to assess primary school children's learning in history, *Cambridge Journal of Education*, 23, pp. 137–154.

HATTIE, J., BIGGS, J. & PURDIE, N. (1996) Effects of learning skills interventions on student learning: a meta-analysis, *Review of Educational Research*, 66, pp. 99–136.

HAYES, S.C., NELSON, R.O. & JARRETT, R.B. (1983) The treatment utility of assessment: a functional approach to evaluating assessment quality, *American Psychologist*, 62, pp. 963–974.

HERMAN, J.L., GEARHART, M. & ASCHBACHER, P.R. (1996) Portfolios for classroom assessment: design and implementation issues, in: R.C. CALFEE & P. PERFUMO (Eds) *Writing Portfolios in the Classroom*, pp. 27–59 (Mahwah, NJ, Lawrence Erlbaum).

HIGGINS, K.M., HARRIS, N.A. & KUEHN, L.L. (1994) Placing assessment into the hands of young children: a study of student-generated criteria and self-assessment, *Educational Assessment*, 2, pp. 309–324.

HOGAN, K. & PRESSLEY, M. (Eds) (1997) *Scaffolding Student Learning: instructional approaches and issues* (Cambridge, MA, Brookline).

HOLLIDAY, W.G. & BENSON, G. (1991) Enhancing learning using questions adjunct to science charts, *Journal of Research in Science Teaching*, 28, pp. 523–535.

HOLLIDAY, W.G. & MCGUIRE, B. (1992) How can comprehension adjunct questions focus students attention and enhance concept-learning of a computer-animated science lesson? *Journal of Research In Science Teaching*, 29, pp. 3–16.

HUGHES, I.E. & LARGE, B.J. (1993) Staff and peer-group assessment of oral communication skills, *Studies in Higher Education*, 18, pp. 379–385.

IREDALE, C. (1990) Pupils' attitudes towards GASP (Graded Assessments in Science Project), *School Science Review*, 72(258) pp. 133–137.

IVERSON, A.M., IVERSON, G.L. & LLUKIN, L.E. (1994) Frequent, ungraded testing as an instructional strategy, *Journal of Experimental Education*, 62, pp. 93–101.

JAMES, M. (1990) Negotiation and dialogue in student assessment, in: T. HORTON (Ed.) *Assessment Debates*, pp. 104–115 (London, Hodder and Stoughton).

JOHNSON, D.W. & JOHNSON, R.T. (1990) Co-operative learning and achievement, in: S. SHARAN (Ed.) *Co-operative Learning: theory and research*, pp. 23–27 (New York, Praeger).

JOHNSTON, P., GUICE, S., BAKER, K., MALONE, J. & MICHELSON, N. (1995) Assessment of teaching and learning in literature-based classrooms, *Teaching and Teacher Education*, 11, pp. 359–371.

KHALAF, A.S.S. & HANNA, G.S. (1992) The impact of classroom testing frequency on high-school-students' achievement, *Contemporary Educational Psychology*, 17, pp. 71–77.

KING, A. (1990) Enhancing peer interaction and learning in the classroom through reciprocal questioning, *American Educational Research Journal*, 27, pp. 664–687.

KING, A. (1991) Effects of training in strategic questioning on children's problem-solving performance, *Journal of Educational Psychology*, 83, pp. 307–317.

KING, A. (1992a) Comparison of self-questioning, summarizing, and note-taking review as strategies for learning from lectures, *American Educational Research Journal*, 29, pp. 303–323.

KING, A. (1992b)/ Facilitating elaborative learning through guided student-generated questioning, *Educational Psychologist*, 27, pp. 111–126.

KING, A. (1994) Autonomy and question asking—the role of personal control in guided student-generated questioning, *Learning and Individual Differences*, 6, pp. 163–185.

KING, A. (1995) Inquiring minds really do want to know—using questioning to teach critical thinking, *Teaching of Psychology*, 22, pp. 13–17.

KING, A. & ROSENSHINE, B. (1993) Effects of guided cooperative questioning on children's knowledge construction, *Journal of Experimental Education*, 61, pp. 127–148.

KLENOWSKI, V. (1995) Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England, *Assessment in Education*, 2, pp. 145–163.

KLUGER, A.N. & DENISI, A. (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory, *Psychological Bulletin*, 119, pp. 254–284.

KOCH, A. & SHULAMITH, G.E. (1991) Improvement of reading comprehension of physics texts by students' question formulation, *International Journal of Science Education*, 13, pp. 473–485.

KRAMER, J.J. (Ed.) (1993) *Curriculum-Based Measurement* (Lincoln, NE, Buros Institute of Mental Measurements).

KULIK, C.-L.C. & KULIK, J.A. (1987) Mastery testing and student learning: a meta-analysis, *Journal of Educational Technology Systems*, 15, pp. 325–345.

KULIK, J.A. & KULIK, C.-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13, pp. 221–340.

KULIK, C.-L.C., KULIK, J.A. & BANGERT-DROWNS, R.L. (1990) Effectiveness of mastery-learning programs: a meta-analysis, *Review of Educational Research*, 60, pp. 265–299.

LAN, W.Y., BRADLEY, L. & PARR, G. (1994) The effects of a self-monitoring process on college students' learning in an introductory statistics course, *Journal of Experimental Education*, 62, pp. 26–40.

LEPPER, M.R. & HODELL, M. (1989) Intrinsic motivation in the classroom, in: C. AMES & R. AMES (Eds) *Research on Motivation in the Classroom*, Vol. 3, pp. 73–105 (San Diego, CA, Academic Press).

LIDZ, C.S. (1995) Dynamic assessment and the legacy of L.S. Vygotsky, *School Psychology International*, 16, pp. 143–153.

LIVINGSTON, J.A. & GENTILE, J.R. (1996) Mastery learning and the decreasing variability hypothesis, *Journal of Educational Research*, 90, pp. 67–74.

LOCK, R. & FERRIMAN, B. (1988) 0CEA—The development of a graded assessment scheme in science Part III School Trials 1986, *School Science Review*, 70(252), pp. 103–112.

LOCK, R. & WHEATLEY, T. (1990) Recording process, skills and criterion asessments—student systems, *School Science Review*, 71(255), pp. 145–150.

LORSBACH, A W., TOBIN, K., BRISCOE, C. & LAMASTER, S.U. (1992) An interpretation of assessment methods in middle school science, *International Journal of Science Education*, 14, pp. 305–317.

LUNDEBERG, M.A. & FOX, P.W. (1991) Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research*, 61, pp. 94–106.

MAQSUD, M. & PILLAI, C.M. (1991) Effect of self-scoring on subsequent performances in academic achievement tests, *Educational Research*, 33, pp. 151–154.

MARTINEZ, J.G.R. & MARTINEZ, N.C. (1992) Re-examining repeated testing and teacher effects in a remedial mathematics course, *British Journal of Educational Psychology*, 62, pp. 356–363.

MASTERS, G.N. & EVANS, J. (1986) A sense of direction in criterion referenced assessment, *Studies in Educational Evaluation*, 12, pp. 257–265.

MAVROMATTIS, Y. (1996) Classroom assessment in Greek primary schools, *The Curriculum Journal*, 7, pp. 259–269.

MCCALLUM, B., MCALISTER, S., GIPPS, C. & BROWN, M. (1993) Teacher assessment at Key Stage 1, *Research Papers in Education*, 8, pp. 305–327.

MCCURDY, B.L. & SHAPIRO, E.S. (1992) A comparison of teacher-monitoring, peer-monitoring, and self-monitoring with curriculum-based measurement in reading among students with learning disabilities, *Journal of Special Education*, 26, pp. 162–180.

MCNEIL, J.D. (1969) Forces influencing curriculum, *Review of Educational Research*, 39, pp. 293–318.

MERRETT, J. & MERRETT, F. (1992) Classroom management for project work: an application for correspondence training, *Educational Studies*, 18, pp. 3–10.

MESSICK, S. (1989) Validity, in: R.L. LINN (Ed.) *Educational Measurement*, 3rd edn, pp. 12–103 (London, Collier Macmillan).

MEYER, K. & WOODRUFF, E. (1997) Consensually driven explanation in science teaching, *Science Education*, 80, pp. 173–192.

MILKENT, M.M. & ROTH, W.-M. (1989) Enhancing student achievement through computer generated homework, *Journal of Research in Science Teaching*, 26, pp. 567–573.

MILLS, R.P. (1996) Statewide Portfolio Assessment: the Vermont Experience, in: J.B. BARON & P. WOLF (Eds) *Performance-Based Student Assessment: challenges and possibilities*, pp. 192–214 (Chicago, IL, University of Chicago Press).

MITCHELL, T. (1991) *Colonising Egypt* (University of California Press).

MORTIMORE, P., SAMMONS, P., STOLL, L. & ECOB, R. (1988) *School Matters: the junior years* (Somerset, Open Books).

MORY, E.H. (1992) The use of informational feedback in instruction—implications for future research, *Educational Technology Research and Development*, 40(3) pp. 5–20.

NATRIELLO, G. (1987) The impact of evaluation processes on students, *Educational Psychologist*, 22, pp. 155–175.

NEWMANN, F.M. (1992) The assessment of discourse in social studies, in: H. BERLAK *et al.* (Eds) *Toward a New Science of Educational Testing and Assessment*, pp. 53–69 (Albany, NY, State University of New York Press).

NEWMAN, R.S. & SCHWAGER, M.T. (1995) Students' help seeking during problem solving: effects of grade, goal, and prior achievement, *American Educational Research Journal*, 32, pp. 352–376.

NICHOLS, P.D. (1994) A framework for developing cognitively diagnostic assessments, *Review of Educational Research*, 64, pp. 575–603.

OTERO, J.C. & CAMPANARIO, J.M. (1990) Comprehension evaluation and regulation in learning from science texts, *Journal of Research in Science Teaching*, 27, pp. 447–460.

PENNYCUICK, D.B. & MURPHY, R.J.L. (1986) The impact of the graded test movement on classroom teaching and learning, *Studies in Educational Evaluation*, 12, pp. 275–279.

PERRENOUD, P. (1991) Towards a pragmatic approach to formative evaluation, in: P. WESTON (Ed.) *Assessment of Pupils' Achievement: Motivation and School Success*, pp. 79–101 (Amsterdam, Swets and Zeitlinger).

PHYE, G.D. (Ed.) (1997) *Handbook of Classroom Assessment* (San Diego CA and London, Academic Press).

PIJL, S.J. (1992) Practices in monitoring student progress, *International Review of Education*, 38, pp. 117–131.

POLLARD, A., BROADFOOT, P., CROLL, P., OSBORN, M. & ABBOTT, D. (1994) *Changing English Primary Schools? The Impact of the Education Reform Act at Key Stage One* (London, Cassell).

POWELL, S.D. & MAKIN, M. (1994) Enabling pupils with learning difficulties to reflect on their own thinking, *British Educational Research Journal*, 20, pp. 579–593.

PRESSLEY, M., WOOD, E., WOLOSHYN, V.E., MARTIN, V., KING, A. & MENKE, D. (1992) Encouraging mindful use of prior knowledge—attempting to construct explanatory answers facilitates learning, *Educational Psychologist*, 27, pp. 91–109.

PRYOR, J. & TORRANCE, H. (1996) Teacher-pupil interaction in formative assessment: assessing the work or protecting the child? *The Curriculum Journal*, 7, pp. 205–226.

PURDIE, N. & HATTIE, J. (1996) Cultural differences in the use of strategies for self-regulated learning, *American Educational Research Journal*, 33, pp. 845–871.

QUICKE, J. & WINTER, C. (1994) Teaching the language of learning: towards a metacognitive approach to pupil empowerment, *British Educational Research Journal*, 20, pp. 429–445.

RADNOR, H.A. (1994) The problems of facilitating qualitative formative assessment in pupils, *British Journal of Educational Psychology*, 64, pp. 145–160.

RAMAPRASAD, A. (1983) On the definition of feedback, *Behavioral Science*, 28, pp. 4–13.

RATCLIFFE, M. (1992) The implementation of criterion-referenced assessment in the teaching of science, *Research in Science and Technological Education*, 10, pp. 171–185.

REYNOLDS, S., MARTIN, K. & GROULX, J. (1995) Patterns of understanding, *Educational Assessment*, 3, pp. 363–371.

RISMARK, M. (1996) The likelihood of success during classroom discourse, *Scandinavian Journal of Educational Research*, 40, pp. 57–68.

RODRIGUES, S. & BELL, B. (1995) Chemically speaking: a description of student-teacher talk during chemistry lessons using and building on students' experiences, *International Journal of Science Education*, 17, pp. 797–809.

ROSENSHINE, B., MEISTER, C. & CHAPMAN, S. (1996) Teaching students to generate questions: a review of the intervention studies, *Review of Educational Research*, 66, pp. 181–221.

ROSS, J.A. (1995) Effects of feedback on student behaviour in cooperative learning groups in a grade-7 math class, *Elementary School Journal*, 96, pp. 125–143.

ROSS, M., RADNOR, H., MITCHELL, S. & BIERTON, C. (1993) Assessing Achievement in the Arts (Buckingham, Open University Press).

ROTH, W-M. & ROYCHOUDHURY, A. (1994) Science discourse through collaborative concept mapping: new perspectives for the teacher, *International Journal of Science Education*, 16, pp. 437–455.

ROWE, K.J. & HILL, P.W. (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles', *Assessment in Education*, 3, pp. 309–351.

RUDMAN, H.C. (1987) Testing and teaching: two sides of the same coin? *Studies in Educational Evaluation*, 13, pp. 73–90.

RUSSELL, T.A., QUALTER, A. & MCGUIGAN, L. (1995) Reflections on the implementation of National Curriculum Science Policy for the 5–14 age range: findings and interpretations from a national evaluation study in England, *International Journal of Science Education*, 17, pp. 481–492.

RYAN, A.G. (1988) Program evaluation within the paradigm: mapping the territory, *Knowledge: creation, diffusion, utilization*, 10, pp. 25–47.

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

SALVIA, J. & HUGHES, C. (1990) *Curriculum Based Assessment* (New York, USA Macmillan).

SALVIA, J. & YSSELDYKE, J.E. (1991) *Assessment* (Boston, MA, Hougton Mifflin).

SAWYER, R.J., GRAHAM, S. & HARRIS, K.R. (1992) Direct teaching, strategy instruction, and strategy instruction with explicit self-regulation: effects on the composition skills and self-efficacy of students with learning disabilities, *Journal of Educational Psychology*, 84, pp. 340–352.

SCHILLING, M., HARGREAVES, L., HARLEN, W. & RUSSELL, T. (1990) *Written Tasks* (London, Paul Chapman).

SCHLOSS, P.J., SMITH, M.A. & POSLUZSNY, M. (1990) The impact of formative and summative assessment upon test performance of science education majors, *Teacher Education and Special Education Majors*, 13, pp. 3–8.

SCHOENFELD, A.H. (1985) *Mathematical problem-solving* (New York, NY, Academic Press).

SCHUNK, D.H. (1996) Goal and self-evaluative influences during children's cognitive skill learning, *American Educational Research Journal*, 33, pp. 359–382.

SCHUNK, D.H. & RICE, J.M. (1991) Learning goals and progress feedback during reading comprehension instruction, *Journal of Reading Behaviour*, 23, pp. 351–364.

SCHUNK, D.H. & SWARTZ, C.W. (1993a) Goals and progress feedback: effects on self-efficacy and writing achievement, *Contemporary Educational Psychology*, 18, pp. 337–354.

SCHUNK, D.H. & SWARTZ, D.W. (1993b) Progress feedback and goals, *Roeper Review*, 15, pp. 225–230.

SCOTT, D. (1991) Issues and themes: coursework and coursework assessment in the GCSE, *Research Papers in Education*, 6, pp. 3–19.

SENK, S.L., BECKMAN, C.E. & THOMPSON, D.R. (1997) Assessment and grading in high school mathematics classrooms, *Journal for Research in Mathematics Education*, 28, pp. 187–215.

SHEPARD, L.A (1995) Using assessment to improve learning, *Educational Leadership*, 52(5), pp. 38–43.

SHEPARD, L.A., FLEXER, R.J., HIEBERT, E.J., MARION, S.F., MAYFIELD, V. & WESTON, T.J. (1994) Effects of introducing classroom performance assessments on student learning, in: *Proceedings of Annual Meeting of AERA Conference*, New Orleans: Available from ERIC ED 390918.

SHEPARD, L.A., FLEXER, R.J., HIEBERT, E.J., MARION, S.F., MAYFIELD, V. & WESTON, T.J. (1996)

Effects of introducing classroom performance assessments on student learning, *Educational Measurement Issues and Practice*, 15, pp. 7–18.

SHINN, M.R. & GOOD III, R.H. (1993) CBA: an assessment of its current status and prognosis for its future, in: J.J. KRAMER (Ed.) *Curriculum-Based Measurement*, pp. 139–178 (Lincoln, NE, Buros Institute of Mental Measurements).

SHINN, M.R. & HUBBARD, D.D. (1992) Curriculum-based measurement and problem-solving assessment—basic procedures and outcomes, *Focus On Exceptional Children*, 24(5), pp. 1–20.

SHOHAMY, E. (1995) Language Testing: matching assessment procedures with language knowledge, in: M. BIRENBAUM & F.J.R.C. DOCHY (Eds) *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*, pp. 143–160 (Boston, Kluwer).

SIERO, F. & VAN OUDENHOVEN, J.P. (1995) The effects of contingent feedback on perceived control and performance, *European Journal of Psychology of Education*, 10, pp. 13–24.

SIMMONS, M. & COPE, P. (1993) Angle and rotation: effects of differing types of feedback on the quality of response, *Educational Studies in Mathematics*, 24, pp. 163–176.

SIMPSON, M. (1990) Why criterion-referenced assessment is unlikely to improve learning, *The Curriculum Journal*, 1, pp. 171–183.

SKAALVIK, E.M. (1990) Attribution of perceived academic results and relations with self-esteem in senior high school students, *Scandinavian Journal of Educational Research*, 34, pp. 259–269.

SLATER, T.F., RYAN, J.M. & SAMSON, S.L. (1997) Impact and dynamics of portfolio assessment and traditional assessment in a college physics course, *Journal of Research in Science Teaching*, 34, pp. 255–271.

SLAVIN, R.E. (1987) Mastery learning reconsidered, *Review of Educational Research*, 57, pp. 175–214.

SLAVIN, R.E. (1991) Synthesis of research on cooperative learning, *Educational Leadership*, 48(5), pp. 71–82.

SLAVIN, R.E., MADDEN, N.A., KARWEIT, N.L., DOLAN, L.J., WASIK, B.A., SHAW, E., MAINZER, K.L., PETZA, R., BOND, M.A. & HAXBY, B. (1992) Success for all: a relentless approach to prevention and early intervention in elementary schools. ERS Monograph (Arlington, VA, Educational Research Service).

SLAVIN, R.E., MADDEN, N.A., DOLAN, L. J., WASIK, B.A., ROSS, S., SMITH, L. & DIANDA, M. (1996) Success for all: a summary of research, *Journal of Education for Students Placed at Risk (JESPAR)*, 1, pp. 41–76.

SOLOMON, J. (1991) Group discussions in the classroom, *School Science Review*, 72(261), pp. 29–34.

STEFANI, L.A.J. (1994) Peer, self and tutor assessment: relative reliabilities, *Studies in Higher Education*, 19, pp. 69–75.

STIGGINS, R.J., GRISWOLD, M.M. & WIKELUND, K.R. (1989) Measuring thinking skills through classroom assessment, *Journal of Educational Measurement*, 26, pp. 233–246.

STRAWITZ, B.M. (1989) The effect of testing on science process skill achievement, *Journal of Research in Science Teaching*, 26, pp. 659–664.

SWAIN, J. (1988) GASP The graded assessments in science project, *School Science Review*, 70(251), pp. 152 –158.

SWAIN, J.R.L. (1989) The development of a framework for the assessment of process skills in a Graded Assessments in Science Project, *International Journal of Science Education*, 11, pp. 251–259.

SWAIN, J.R.L. (1991) The nature and assessment of scientific explorations in the classroom, *School Science Review*, 72(260), pp. 65–76.

TAN, C.M. (1992) An evaluation of the use of continuous assessment in the teaching of physiology, *Higher Education*, 23, pp. 255–272.

TENENBAUM, G. & GOLDRING, E. (1989) A meta-analysis of the effect of enhanced instruction: cues, participation, reinforcement and feedback and correctives on motor skill learning, *Journal of Research and Development in Education*, 22(3), pp. 53–64.

THOMAS, J.W. (1993) Promoting independent learning in the middle grades—the role of instructional support practices, *Elementary School Journal*, 93, pp. 575–591.

THOMAS, J.W., BOL, L., WARKENTIN, R.W., WILSON, M., STRAGE, A. & ROHWER, W.D. (1993) Interrelationships among students' study activities, self-concept of academic ability, and achievement as a function of characteristics of high-school biology courses, *Applied Cognitive Psychology*, 7, pp. 499–532.

TINDAL, G. (1993) A review of curriculum-based procedures on nine assessment components, in: J.J. KRAMER (Ed.) *Curriculum-Based Measurement*, pp. 25–64 (Lincoln, NE, Buros Institute of Mental Measurements).

TITTLE, C.K. (1994) Toward an educational-psychology of assessment for teaching and learning—theories, contexts, and validation arguments, *Educational Psychologist*, 29, pp. 149–162.

TORRANCE, H. (1993) Formative assessment: some theoretical problems and empirical questions, *Cambridge Journal of Education*, 23, pp. 333–343.

TORRANCE, H. & PRYOR, J. (1995) Investigating teacher assessment in infant classrooms: methodological problems and emerging issues, *Assessment in Education*, 2, pp. 305–320.

TORRIE, I. (1989) Developing achievement based assessment using grade related criteria; *Research in Science Education*, 19, pp. 286–290.

TUNSTALL, P. & GIPPS, C. (1996a) 'How does your teacher help you to make your work better?' Children's understanding of formative assessment, *The Curriculum Journal*, 7, pp. 185–203.

TUNSTALL, P. & GIPPS, C. (1996b) Teacher feedback to young children in formative assessment: a typology, *British Educational Research Journal*, 22, pp. 389–404.

VISPOEL, W.P. & AUSTIN, J.R. (1995) Success and failure in junior high school: a critical incident approach to understanding students' attributional beliefs, *American Educational Research Journal*, 32, pp. 377–412.

VON GLASERSFELD, E. (1987) Learning as a constructive activity, in: C. JANVIER (Ed.) *Problems of Representation in the Teaching and Learning of Mathematics* (Hillsdale, NJ, Lawrence Erlbaum Associates).

WEBB, N.M. (1995) Group collaboration in assessment: multiple objectives, processes, and outcomes, *Educational Evaluation and Policy Analysis*, 17, pp. 239–261.

WESTON, C., MCALPINE, L. & BORDONARO, T. (1995) A model for understanding formative evaluation in instructional design, *Educational Technology Research and Development*, 43(3), pp. 29–48.

WHELDALL, K. & PANGAGOPOLOU-STEAMATELATOU, A. (1992) The effects of pupil self-recording of on-task behaviour on primary school children, *British Educational Research Journal*, 17, pp. 113–127.

WHITING, B., VAN BURGH, J.W. & RENDER, G.F. (1995) Mastery learning in the classroom, paper presented at the *Annual Meeting of the AERA San Francisco 1995*, available from ERIC ED382688.

WILIAM, D. (1994) Assessing authentic tasks: alternatives to mark-schemes, *Nordic Studies in Mathematics Education*, 2, pp. 48–68.

WILIAM, D. & BLACK, P.J. (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, 22, pp. 537–548.

WILIAM, D. (1997, February) *Evaluation of the Key Stage 3 Diagnostic Software pilot*, Report prepared for School Curriculum and Assessment Authority (London, UK, King's College London School of Education).

WITHERS, G. (1987) From marking strategy to assessment procedure: a review of recent Australian practices, *Studies in Educational Evaluation*, 13, pp. 7–19.

WOLF, D., BIXBY, J., GLENN, J. & GARDNER, H. (Eds) (1991) *To Use Their Minds Well: investigating new forms of student assessment* (Washington, DC, American Educational Research Association).

YANCEY, K.B. (1996) Dialogue, interplay and discovery: mapping the role and rhetoric of reflection in portfolio assessment, in: R. CALFEE & P. PERFUMO (Eds) *Writing Portfolios in the Classroom*, pp. 83–102 (Mahwah, NJ, Lawrence Erlbaum).

ZESSOULES, R. & GARDNER, H. (1991) Authentic assessment: beyond the buzzword and into the

classroom, in: V. PERRONE (Ed.) *Expanding Student Assessment*, pp. 47–71 (Alexandria Virginia, USA, Association for Supervision and Curriculum Development).

ZIMMERMAN, B.J. & BAMDURA, A. (1994) Impact of self-regulatory influences on writing course attainment, *American Educational Research Journal*, 31, pp. 845–862.

ZIMMERMAN, B.J. & RISEMBERG, R. (1997) Becoming a self-regulated writer: a social cognitive perspective, *Contemporary Educational Psychology*, 22, pp. 73–101.

Appendix: list of journals searched

For the journals listed below, the contents lists were scanned to identify relevant articles, as described in the Introduction. Articles from journals not in this list were found by other means.

1. *American Educational Research Journal*
2. *Applied Cognitive Psychology*
3. *Assessment and Evaluation in Higher Education*
4. *Assessment in Education: principles, policy and practice*
5. *British Educational Research Journal*
6. *British Journal of Curriculum and Assessment*
7. *British Journal of Educational Psychology*
8. *British Journal of Educational Studies*
9. *British Journal of Educational Technology*
10. *Cambridge Journal of Education*
11. *Child Development*
12. *College Teaching*
13. *Contemporary Educational Psychology*
14. *Education Review* (from 91)
15. *Education Technology Research and Development*
16. *Educational and Psychological Measurement*
17. *Educational Assessment*
18. *Educational Evaluation and Policy Analysis*
19. *Educational Leadership*
20. *Educational Measurement: issues and practice*
21. *Educational Psychologist*
22. *Educational Research*
23. *Educational Researcher* (from 92)
24. *Educational Studies in Mathematics*
25. *Educational Technology and Training International*
26. *Elementary School Journal*
27. *English in Education*
28. *European Education*
29. *European Journal of Education*
30. *European Journal of Psychology of Education*
31. *Focus on Exceptional Children*
32. *Harvard Educational Review*
33. *International Journal of Educational Research* (92–96)
34. *International Journal of Mathematics Education*
35. *International Journal of Nursing Studies*
36. *International Journal of Science Education*
37. *International Review of Education*
38. *Journal for Research in Mathematics Education*
39. *Journal of Education* (Boston)
40. *Journal of Educational Measurement*
41. *Journal of Educational Psychology*

42. *Journal of Educational Research*
43. *Journal of Educational and Behavioural Statistics*
44. *Journal of Experimental Education* (up to 94/5)
45. *Journal of Higher Education*
46. *Journal of Personal and Social Psychology* (up to 93)
47. *Journal of Reading*
48. *Journal of Reading Behaviour* (up to 95)
49. *Journal of Research and Development in Education*
50. *Journal of Research in Science Teaching*
51. *Journal of School Psychology*
52. *Journal of Special Education*
53. *Learning and Individual Differences* (from 91)
54. *Learning Disability in Focus* (to 90)
55. *Measurement Issues and Educational Practice* (from 92)
56. *Organizational Behaviour and Human Decision Processes*
57. *Oxford Review of Education*
58. *Phi Delta Kappan*
59. *Psychological Bulletin*
60. *Psychology in the Schools*
61. *Remedial and Special Education*
62. *Research in Science Education*
63. *Research Papers in Education*
64. *Review of Educational Research*
65. *Review of Research in Education*
66. *Scandinavian Journal of Educational Research*
67. *School Psychology International*
68. *Science Education*
69. *Studies in Educational Evaluation*
70. *Studies in Higher Education*
71. *Studies in Science Education*
72. *Teachers College Record*
73. *Teaching and Teaching Education*
74. *The Curriculum Journal*
75. *Westminster Studies in Education*
76. *Young Children*