Review of Educational Research
Winter 1988, Vol. 58, No. 4, pp. 438–481

The Impact of Classroom Evaluation Practices on Students

Terence J. Crooks
University of Otago

In most educational programs, a substantial proportion of teacher and student time is devoted to activities which involve (or lead directly to) evaluation by the teacher of student products or behavior. This review summarizes results from 14 specific fields of research that cast light on the relationships between classroom evaluation practices and student outcomes. Particular attention is given to outcomes involving learning strategies, motivation, and achievement. Where possible, mechanisms are suggested that could account for the reported effects. The conclusions derived from the individual fields are then merged to produce an integrated summary with clear implications for effective educational practice. The primary conclusion is that classroom evaluation has powerful direct and indirect impacts, which may be positive or negative, and thus deserves very thoughtful planning and implementation.

> We say we want sensitive, thoughtful, analytic, independent scholars, then treat them like Belgian geese being stuffed for pâté de foie gras. We reward them for compliance, rather than independence; for giving the answers we have taught themrather than for challenging the conclusions we have reached; for admiring the brilliance of purely scientific advances rather than developing greater sensitivity to the inequities in health care we have too often ignored. As one of my associates has observed, 'We put our medical students in jail for years in order that they shall learn to become free men.' (Miller, 1978)

There has been extensive research on the impact of standardized testing on students, and this research has been repeatedly reviewed (Goslin, 1967; Kellaghan, Madaus, & Airasian, 1982; Kirkland, 1971; Madaus & Airasian, 1977; Madaus & McDonagh, 1979; Rudman et al., 1980). Although standardized tests do have important and widespread effects under some circumstances (such as when students must reach given standards to graduate from high school, or when the funding of school districts is affected by test results), students spend vastly greater amounts of time engaged in classroom evaluation activities than in standardized testing. Further, surveys of teachers and students have consistently indicated that they believe the educational and psychological effects of classroom evaluation are generally substantially greater than the corresponding effects of standardized testing (Dorr-

The author acknowledges support from the University of Otago, University of Michigan, University of Iowa, and the American College Testing Program during the writing of this paper. Some parts of the paper were presented at the 14th International Conference of the International Association for Educational Assessment, Iowa City, IA, May 23–27, 1988. Mina Crooks, Janet Kane, Michael Kane, Wilbert McKeachie and three anonymous reviewers provided very helpful comments on drafts of this paper.

Bremme & Herman, 1986; Haertel, 1986; Kellaghan et al., 1982; Salmon-Cox, 1981; Stiggins & Bridgeford, 1985).

Because classroom evaluation activities appear to have very significant effects on students, this review will synthesize research that relates to the impact of classroom evaluation on students. Research evidence from a wide variety of research domains will be reviewed and summarized, and the conclusions from these domains will be drawn together to identify implications for effective educational practice.

For the purposes of this review, classroom evaluation is defined as evaluation based on activities that students undertake as an integral part of the educational programs in which they are enrolled. These activities may involve time spent both inside and outside the classroom. This definition includes tasks such as formal teacher-made tests, curriculum-embedded tests (including adjunct questions and other exercises intended to be an integral part of learning materials), oral questions asked of students, and a wide variety of other performance activities (cognitive and psychomotor). It also includes assessment of motivational and attitudinal variables and of learning skills.

Formal testing under carefully controlled conditions is often only a small component of the total set of evaluation activities in a course (especially in the early years of schooling), but the impact of classroom testing on students has been studied much more extensively than the impact of other forms of classroom evaluation. Thus tests and test-like activities feature prominently in this review. Other forms of classroom evaluation undoubtedly also have important effects on students. Fortunately, many of the general conclusions that can be drawn from research on testing are likely to apply also to other forms of classroom evaluation.

I have chosen to discuss research that was conducted in laboratory settings, even though it may seem to have little ecological validity for classroom evaluation. Much of the classroom-based research also has very limited ecological validity, due to artificial experimental conditions, very brief treatments, or other factors. The application of almost all educational research to new settings or conditions requires thoughtful analysis and sensitivity to factors that may affect the relevance or applicability of the findings in the new settings, or with particular categories of people. As Cronbach (1975) has put it,

> Systematic inquiry can reasonably hope to make two contributions. One reasonable aspiration is to assess local events accurately, to improve short-run control. The other reasonable aspiration is to develop explanatory concepts, concepts that will help people use their heads. (p. 126)

Although generalizations must be made with extreme caution (Cronbach, 1975; Crooks, 1982; McKeachie, 1974), it seems desirable to look for parallels or convergence between the broad findings of research in different domains. Thus, for instance, the present review finds that research in three quite different domains (teacher oral questions, adjunct questions in reading passages, and teacher-made tests) that examined the effects of using questions of higher or lower cognitive levels seems to converge quite nicely. This enhances confidence in the generality of the findings.

The review is structured in four sections. The first section takes a preliminary look at the nature, role, and impact of classroom evaluation. The second section

reviews research that focuses primarily on the impact of various classroom evaluation practices on student learning activities and achievement. Nine specific areas of research are included. The third section reviews research on student motivation, examining the effects of different evaluation practices on motivation, and the consequences of resulting motivational tendencies for student learning. Five areas of motivational research are included. The final section draws together the major findings from the second and third sections and indicates the implications of these findings for the effective use of classroom evaluation in education.

The Nature, Role, and Impact of Classroom Evaluation: An Overview

This section includes three subsections. The first briefly summarizes the findings of research on existing patterns of classroom evaluation in elementary and secondary schools. The second discusses and categorizes the variables that are assessed through classroom evaluation. The third lists 17 specific ways (categorized as short, medium, or long term) in which classroom evaluation affects students.

Patterns of classroom evaluation. In the past few years, several research teams and individuals have examined classroom evaluation practices in elementary, junior high, and high schools in some detail (Dorr-Bremme & Herman, 1986; Fennessy, 1982; Fleming & Chambers, 1983; Gullickson, 1984, 1985; Gullickson & Ellwein, 1985; Haertel, 1986; Stiggins, 1985; Stiggins & Bridgeford, 1985; Stiggins, Conklin, & Bridgeford, 1986). Their findings are summarized below.

A substantial proportion of student time is involved in activities that are evaluated. In two studies (Dorr-Bremme & Herman, 1986; Haertel, 1986), tests occupied students for 5 to 15% of their time on average, with the lower figure being more typical for elementary school students and the higher figure for high school students. However, this was only the time spent on taking formal written tests. Much additional time is spent on other activities that are evaluated, formally or informally. Particular emphasis is placed on these nontest approaches at the elementary level (Gullickson, 1985).

A wide range of evaluative activities takes place in classrooms, with the pattern varying markedly at different grade levels and in different subject areas (Fennessy, 1982; Gullickson, 1985; Stiggins & Bridgeford, 1985). Activities include evaluation through teacher questioning and class or group discussion, marking or commenting on performances of various kinds, checklists, informal observation of learning activities, teacher-made written tests, and written exercises of various kinds (including projects, assignments, worksheets, text-embedded questions, and tests). Affective variables (e.g., aspects of motivation) are also assessed, usually in informal ways.

Teachers judge evaluative activities to be important aspects of teaching and learning and work at them accordingly, but are often concerned about the perceived inadequacies in their efforts (Gullickson, 1984; Stiggins & Bridgeford, 1985).

A substantial proportion of teachers have little or no formal training in educational measurement techniques, and many of those who do have such training find it of little relevance to their classroom evaluation activities (Gullickson, 1984; Gullickson & Ellwein, 1985; Haertel, 1986; Stiggins, 1985). This is especially true for elementary school teachers because of their heavy reliance on observation and other nontest means of evaluation. There are strong arguments for helping teachers to improve these nontest forms of evaluation (e.g., Shulman, 1980, pp. 69–70).

What is evaluated? Bloom (1956) classified educational outcomes into three major domains: cognitive, affective, and psychomotor; and subdivided the cognitive domain into the six well known categories (knowledge, comprehension, application, analysis, synthesis, and evaluation). Other researchers have used different classification schemes. For instance, Gagne, Briggs, and Wager (1988) identified five categories of learned outcomes: intellectual skills, cognitive strategies, verbal information, attitudes, and motor skills. Their first three categories could be seen as a subdivision of Bloom's cognitive domain, although the category labeled cognitive strategies was not directly addressed in the Bloom taxonomy. Most educators would agree that objectives in all three domains are important outcomes of education, with the relative importance of the different domains varying somewhat by subject area.

This review focuses primarily on research examining effects of evaluation on the cognitive domain and certain aspects of the affective domain (test anxiety, student self-efficacy, intrinsic motivation, attributions for success and failure, and cooperation among students). This is not intended to imply the unimportance of other learning outcomes, and I would expect that most of the effects identified in this review would have close parallels in other areas. Also, increasing attention is being given to evaluation of the processes teachers and students use in the pursuit of learning outcomes. This is certainly an area that should not be neglected: Many studies reviewed here demonstrate that the learning strategies students adopt are powerful predictors of educational outcomes, so that expertise in the selection and application of learning strategies is an important educational outcome.

Many researchers and educators have found Bloom's six level taxonomy of the cognitive domain difficult to apply in practice. Any given test question may be answered in different ways by different students, depending on their specific past experiences with the topic (see, for instance, the discussion by Haertel, 1985, p. 33, of research by Nuthall and Lee). Although these difficulties do not fully vanish when levels of the taxonomy are collapsed, many authors have found it more satisfactory to use a simplified version of the taxonomy with two or three levels. The most common approach is to retain knowledge (recall or recognition of specific information) as one category and to form two levels from the remaining categories (see, for example, Buckwalter, Schumacher, Albright, & Cooper, 1981; Crooks & Collins, 1986; Mathews, 1980; and Rinchuse & Zullo, 1986). The second level usually includes Bloom's comprehension category and the capacity to perform routine, well-practiced application of knowledge. The third level is often described as problem solving, the key feature of which is the transfer of existing knowledge and skills to situations that the student has not met before.

Of course, for some purposes specially designed taxonomies are desirable. For instance, in view of the heavy emphasis on Bloom's knowledge level in teacher-made tests in schools, when analyzing 342 of these tests, Fleming and Chambers (1983) broke the knowledge level into three sublevels. Likewise, within the specific domain of problem solving, Fredericksen (1984b) has identified three classes of problems: well structured problems, structured problems requiring productive thinking, and ill structured problems. The first of these, involving routine procedures such as calculating the area of a right triangle, would not even be classified as problem solving in the three level taxonomy described in the last paragraph, but Fredericksen's label for this type of activity is consistent with the widespread use

of description, in textbooks and courses, of routine exercises as "problems." At the other extreme, as Fredericksen and others have noted, many real problems are ill structured, but such problems are often avoided in our education systems because of their complexity and open-endedness.

Such terms as *higher level* questions, *thinking skills*, and *problem solving* are widely used in the research summarized here. In light of the discussion above, however, it is not surprising that there is much inconsistency in the way these terms have been defined (see, for instance, Carrier & Fautsch-Partridge, 1981, for a discussion of categories used in one research area). Careful attention to the particular definitions used in each research report is thus essential.

Several researchers have used coding schemes to analyze the cognitive levels of questions included in teacher-made tests, at grade levels ranging from elementary school to university (Ball et al., 1986; Black, 1968; Buckwalter et al., 1981; Crooks & Collins, 1986; Fleming & Chambers, 1983; Haertel, 1986; Milton, 1982; Rinchuse & Zullo, 1986; Stiggins, Griswold, Green, & associates, 1988). In general, these studies have revealed extensive use of questions at Bloom's lowest ("knowledge") level. For instance, after analyzing 8800 test questions from tests in 12 grade and subject area combinations (elementary to high school), Fleming and Chambers (1983) reported that almost 80% of all questions were at the knowledge level. Mathematics and French contributed most of the higher level items. Similarly, Haertel (1986) found that "classroom examinations often failed to reflect teachers' stated instructional objectives, frequently requiring little more than repetition of material presented in the textbook or class, or solution of problems much like those encountered during instruction" (p. 2).

This finding is not unexpected. Indeed, both proponents and critics of educational testing widely agree that teacher-made tests tend to give greater emphasis to lower cognitive levels than the teachers' stated objectives would justify. Several possible causes have been suggested. These include the difficulty of writing items (especially the widely used short answer and objective items) to assess comprehension and higher level skills, the greater ease with which teachers can defend their marking of questions involving recall or recognition and achieve tests with high reliability (Elton, 1982, pp. 115–116; Natriello, 1987, p. 158), and the belief of teachers that the use of higher level questions will result in confusion, anxiety, and significant levels of failure (Doyle, 1983, 1986). Nevertheless, this pattern is a cause for concern, both because it reduces the validity of teacher's evaluations of their students and because this review will present strong evidence that the use of higher level questions in evaluation enhances learning, retention, transfer, interest, and development of learning skills.

Other aspects of the case for reduced emphasis on testing recall and recognition of factual knowledge are presented by Broudy (1988), Cole (1986), DiSibio (1982), Ebel (1982), Glaser (1985), Linn (1983), Messick (1984a, 1984b), Quellmaltz (1985), Rothkopf (1988), and Thorndike (1969). While they differ in focus and emphasis, they tend to agree that transfer is a very important quality of learning. Thorndike puts it particularly well:

> The crucial indicator of a student's understanding of a concept, a principle, or a procedure is that he is able to apply it in circumstances that are different from those under which it was taught. Transferability is the key feature of meaningful

> learning. So if we are to test for understanding, we must test in circumstances which are at least part new. (Thorndike, 1969, p. 2)

Clearly, educational achievement must be seen as substantially more than the accumulation of isolated pieces of information and the development of certain overlearned skills that can be reliably performed. Indeed, Broudy (1988) argues that neither the *replicative* nor the *applicative* uses of schooling are sufficient to make a good case for general education of the whole school population. Rather, he argues, one must look to what he calls the *associative* and *interpretive* uses of schooling to build such a case.

Ways in which evaluations affect students: An overview. Evaluations affect students in short, medium, and long term ways. I have classified the effects into three groups based on this time perspective. There are inevitably some parallels between effects in the different categories.

At the level of a particular lesson, topic, or assignment, the following effects seem to apply (see Gagne, 1977, for a similar list):

1. Reactivating or consolidating prerequisite skills or knowledge prior to introducing the new material;
2. Focusing attention on important aspects of the subject;
3. Encouraging active learning strategies;
4. Giving students opportunities to practice skills and consolidate learning;
5. Providing knowledge of results and corrective feedback;
6. Helping students to monitor their own progress and develop skills of self-evaluation;
7. Guiding the choice of further instructional or learning activities to increase mastery;
8. Helping students feel a sense of accomplishment.

At the level of a particular learning module, course, or extended learning experience, the following are important effects:

1. Checking that students have adequate prerequisite skills and knowledge to effectively learn the material to be covered;
2. Influencing students' motivation to study the subject and their perceptions of their capabilities in the subject;
3. Communicating and reinforcing (or in some cases undermining) the instructor's or the curriculum's broad goals for students, including the desired standards of performance;
4. Influencing students' choice of (and development of) learning strategies and study patterns;
5. Describing or certifying students' achievements in the course, thus influencing their future activities.

Finally, evaluation has longer term consequences, especially when students meet consistent patterns of evaluation year after year. These longer term effects include:

1. Influencing students' ability to retain and apply in varied contexts and ways the material learned;
2. Influencing the development of students' learning skills and styles;
3. Influencing students' continuing motivation, both in particular subjects and more generally;

4. Influencing the students' self-perceptions, such as their perceptions of their self-efficacy as learners.

These effects have been listed very concisely here, but most of them will be discussed in considerable depth in the next two sections of this paper.

The Impact of Classroom Evaluation on Student Learning Activities and Achievement

This section consists of nine subsections, arranged in two groups. Each subsection presents a brief review of a particular field of research on classroom evaluation practices. Although motivational factors help explain some of the reported findings, and some of the evaluation arrangements discussed have marked effects on motivational and affective outcomes, the prime emphasis in this section is on how the implementation of classroom evaluation affects learning strategies and cognitive outcomes. Motivational influences and outcomes are more fully discussed in the next major section of this review.

The Impact of Normal Classroom Testing Practices

Effects related to expectations of what will be tested: The studying and learning practices of college students. Intensive research on the studying and learning approaches of college students over the past 20 years has identified consistent patterns in the learning strategies adopted by university students and in the relationships between these strategies and teaching arrangements (notably the evaluation approaches used). Although this research has been conducted with college students, the findings seem to have much wider application.

The research has been characterized by extensive use of interviews with students, although later researchers have developed questionnaires to gather data more economically. This research began in the United States with the sociological work of Becker, Geer, and Hughes (1968) and the insightful psychological investigations of the intellectual development of Harvard University students conducted by Perry (1970). Most of the more recent work, however, has been carried out in Europe and Australia. Marton, Säljö, and their colleagues in Sweden gave great impetus to this field with their work in the 1970s, and were the first to identify the patterns that have been verified repeatedly since then (although it should be noted that Perry's earlier work is highly related). This work has been extensively reviewed by Entwistle and Ramsden (1983), Ford (1981), Marton, Hounsell, and Entwistle (1984), Schmeck (1983), and Wilson (1981).

Marton and Säljö (1976a) reported that students' approaches to learning tasks could be categorized into two broad categories that they labeled as *deep* or *surface* approaches. Deep approaches involved an active search for meaning, underlying principles, structures that linked different concepts or ideas together, and widely applicable techniques. Surface approaches, in contrast, relied primarily on attempts to memorize course material, treating the material as if different facts and topics were unrelated. Similar categories have been found in many later studies (see Biggs, 1978; Entwistle & Ramsden, 1983; Marton, Hounsell, & Entwistle, 1984; Ramsden, 1985; and Watkins, 1984), although some researchers have identified subcategories within the surface and deep approaches (van Rossum, Deijkers, & Hamer, 1985).

After the initial study, fóllow-up studies by Marton and Säljö (1976b), Svensson

(1977), Dahlgren (1978), Laurillard (1979), and Ramsden and Entwistle (1981) demonstrated that most students were somewhat versatile in their choice of learning approach. Their choice depended on such factors as their interest in the topic, the nature of their academic motivation, the pressure of other demands on their time and energy, the total amount of content in the course, the way in which a task is introduced, and their perceptions of what will be demanded of them in subsequent evaluations or applications of the material (Entwistle & Ramsden, 1983; Laurillard, 1984; Ramsden, 1985).

The choice of evaluation approaches seemed to be particularly potent in its effect, leading Elton and Laurillard (1979) to conclude that perhaps "here is something approaching a law of learning behaviour for students: namely that the quickest way to change student learning is to change the assessment system" (p. 100).

The effects of evaluation on the studying and learning approaches adopted by students can be positive or negative. Fredericksen (1984a) described these effects as "the real test bias," and illustrated his case with numerous examples from the research literature. More informally, but no less powerfully, Rogers (1969), receiving an award for career contributions to the teaching of physics, described the effects of examinations on students as follows:

> Examinations tell them our real aims, at least so they believe. If we stress clear understanding and aim at a growing knowledge of physics, we may completely sabotage our teaching by a final examination that asks for numbers to be put into memorized formulas. However loud our sermons, however intriguing the experiments, students will judge by that examination—and so will next year's students who hear about it (p. 956).

Some of the qualitative influences of evaluation on learning have been investigated and described in books by Becker et al. (1968), Miller and Parlett (1974), and Snyder (1971). They found that many students aimed to plan their study with the primary goal of performing well on course examinations and other evaluation tasks. Unfortunately, the students often saw this goal as conflicting with the more fundamental goal of gaining a deep and enduring grasp of the subject. At the Massachusetts Institute of Technology, Snyder (1971) found that while what he called the formal curriculum emphasized a problem-oriented approach, originality, and independence of thought, the evaluation (which he called the *hidden curriculum*) tended to emphasize an answer-oriented approach and rote learning. Some students with high intrinsic motivation chose not to let the evaluation system distort their learning goals (for example, the student quoted in Snyder, 1971, p. 36), but the majority were happy to focus mainly on the demands of the evaluation system.

Of course, students differ markedly in their capacity to clearly identify the nature and substance of those demands. Some (Miller & Parlett, 1974, call them *cue seekers*) are very adept and energetic in figuring out optimum strategies for obtaining high marks economically, while others (*cue conscious*) are less active, but take careful note of any cues that come their way, and a minority are *cue deaf*.

Even when students correctly identify this hidden curriculum, they may not be capable of adapting to its demands. Several studies (Martin & Ramsden, 1987; Marton & Säljö, 1976b; Ramsden, 1984; van Rossum & Schenk, 1984) have shown

that students who generally use surface approaches have great difficulty adaptingto evaluation requirements that favor deep approaches. On the other hand, these and other studies have demonstrated that students who on some occasions successfully use deep approaches can all too easily be persuaded to adopt surface approaches if evaluation or other factors suggest that these will be successful. For instance, if an examination consists entirely of detailed factual questions on lecture material, an effective strategy would be to attend all lectures, take detailed notes, and rely on last-minute cramming of the lecture notes in the days immediately before the examination (Crooks & Mahalski, 1986). Miller and Parlett (1974, p. 107) have suggested that such examinations may actually serve to clear from the student's memory the knowledge involved, rather than to strengthen it. Other research suggests this is unlikely, but certainly there is ample research to indicate that detailed factual knowledge decays rapidly unless it is used or restudied.

One interesting illustration of an apparent influence of curriculum and evaluation practices on students emerges from a study by Entwistle and Kozeki (1985). They examined the school motivation, approaches to studying, and attainment of high school students in Britain and Hungary. Using Entwistle's well-established *Approaches to Studying Inventory*, they identified substantial mean differences between British and Hungarian students on the deep and reproducing (surface) approach scales. Compared to the British students, the Hungarian students had higher scores on deep approach and lower scores on surface approach. They convincingly hypothesized that this reflected differences in teaching and examining in the two countries. As they interpreted it, the external examinations in Britain in the latter years of high school place a very heavy emphasis on the correct reproduction of information, and this influences the approaches adopted by both teachers and students. In Hungary, on the other hand, there has been a strong reaction against a former stress on rote learning in the schools, and the emphasis has recently been placed on attempting to foster creativity through helping students to think about relationships, with much reduced emphasis on factual knowledge or operation learning. If Entwistle and Kozeki's interpretation is correct, their findings are a vivid demonstration of the influence of what is emphasized and assessed in school on how students approach their learning.

On a smaller scale, Newble and Jaeger (1983) described the effects of a change in evaluation on students in a medical school. When ward ratings replaced an oral clinical examination, students found that ward ratings were almost always above the pass level. Given that their written theory examinations *did* produce failures, they started spending more time in the library and less in the wards. Instituting a different clinical examination shifted the balance back. Newble and Jaeger commented that the effect of the change was so great as to indicate that examinations may be the major factor influencing student learning in a medical school with a traditional curriculum. A number of similar examples are given by Milton (1982), in a book which critically analyzes college evaluation practices.

Ramsden, Beswick, and Bowden (1987) gave university students training intended to improve their learning skills, expecting the students to make more use of deep approaches as a result. They found, however, the training actually led to an increase in use of surface approaches, because the training had made students more able to analyze the demands of their course evaluation procedures, which suited surface approaches (see also the comments of Schmeck, 1988, p. 180).

All these examples serve to demonstrate that evaluation approaches exert a powerful influence on how students go about their studying. Ericksen (1983), reflecting on a lifetime of research and writing on teaching and learning, left no doubt about one of his conclusions:

> An examination is a revealing statement by a teacher about what is important in the course. In fact, faculty standards concerning A-grade performance may be the most significant single means by which teachers set the academic values of a college. (p. 135)

Thus far in this section, I have reported on research that has demonstrated that evaluation of students often has a major impact on how they go about their studying. However, many of the studies also looked, qualitatively or quantitatively, at the outcomes achieved. These studies have shown that the nature of students' recall of the content is highly related to the strategies used earlier in studying it (e.g., Marton & Säljö, 1976a, 1976b; van Rossum et al., 1985; van Rossum & Schenk, 1984). They have also shown that students adopting deep approaches perform well on the evaluations associated with their courses, apparently doing at least as well on lower cognitive level questions as their surface-oriented peers, and doing much better than those peers on questions at higher levels (Biggs, 1973; Martin & Ramsden, 1987; Svensson, 1977).

Stice (1987), in a vivid autobiographical account of his own academic experiences, describes how he achieved excellent grades through high school and his first 2 years of college by relying solely on surface strategies. By the first year of graduate school, however, these strategies were not producing satisfactory results. With the help of a friend, he painfully developed new and deeper strategies, and this ultimately led to success in graduate school and a distinguished academic career.

In light of the research reviewed in this section, there seems to be a strong case for encouraging the development of deep strategies from the early years of the education system. This would be facilitated by greater emphasis on higher level questions in evaluations of student progress.

Effects relating to expectations of the evaluation format. A substantial number of studies over the past 50 years have examined effects on study behavior and test performance of student expectations (sets) relating to the types of test items they expect to have to answer (e.g., d'Ydewalle, Swerts, & De Corte, 1983; Gay, 1980; Hakstian, 1971; Hunkins, 1969; Kulhavy, Dyer, & Silver, 1975; Kumar, Rabinsky, & Pandey, 1979; Meyer, 1934, 1935; Rickards & Friedman, 1978; Sax & Collet, 1968; Terry, 1933). Unfortunately, synthesis of the results of this research is severely constrained by inconsistency or inadequacy in design of the studies. In several studies, students were told to expect either essay or multiple choice test formats, but were not apparently given examples of the items or practice on similar items. Thus their expectations may have been very general or unclear, and they may have received little guidance as to the cognitive levels covered by the two types of items. Under these circumstances one would not expect strong effects. In other studies, more careful attention was given to establishing clear expectations, but the cognitive levels of practice and criterion items were not reported, leading to difficulties in interpreting the findings.

Where the cognitive level (or range of cognitive levels) of items was similar in

the two groups, with only the item format differing, differences between the groups were generally small (e.g., Hakstian, 1971, who used a mix of cognitive levels, and Kumar et al., 1979, who used only factual questions). Where statistically significant differences were found under these circumstances (e.g., d'Ydewalle et al., 1983; Meyer, 1935, 1936), they favored the group which prepared for a recall (as opposed to a recognition) task. The recall group tended to prepare more thoroughly and perform a little better.

On the whole, though, student expectations of the cognitive level and content of tasks probably exert much more influence on their study behavior and achievement than do their expectations of the task format (for given content and cognitive level). Thus I believe that there is no strong evidence from this research to support widespread adoption of any one item format or style of task. Instead, the basis for selecting item formats should be their suitability for testing the skills and content that are to be evaluated.

A few studies have examined the comparative merits of open book and closed book testing (see Boniface, 1985; Francis, 1982). These studies have shown that students tend to be less anxious about open book tests, and to prepare somewhat less thoroughly for them. Predictably, the students who rely most on using their notes and/or textbooks during the test tend to be among the lower achievers. Studies to date have demonstrated no clear benefit in levels of student achievement arising from open book tests. More research is needed, however, because most of the treatments have been very brief and thus have not allowed the students adequate opportunity to develop skills in handling the demands of open book tests. Also, more attention to the nature of the test is needed because the availability of resource materials is most likely to be meaningful and useful when tests are not speeded and consist of higher cognitive level questions.

Effects of frequency of testing. The substantial body of research on the effects on students of the frequency of classroom testing has been thoroughly reviewed in a meta-analysis by Bangert-Drowns, Kulik, and Kulik (1988). This review will draw heavily on their work.

The review by Bangert-Drowns et al. (1988) used data from 31 studies which: (a) were conducted in real classrooms, (b) had all groups receiving the same instruction except for varying frequencies of testing, (c) used conventional classroom tests, (d) did not have serious methodological flaws, and (e) used a summative end-of-course examination taken by all groups as a dependent variable. The course length ranged from 4 weeks to 18 weeks, but only 9 studies were of courses shorter than 10 weeks. Bangert-Drowns et al. reported their results in terms of effect size (difference in mean scores divided by standard deviation of the less frequently tested group).

Overall, they found an effect size of 0.25 favoring the frequently tested group, representing a modest gain in examination performance associated with frequent testing. However, the actual frequencies of testing varied dramatically, so the collection of studies was very heterogeneous. In 12 studies where the low frequency group received no testing prior to the summative examination, the effect size increased to 0.43. It seems reasonable to hypothesize that, in part, this large increase may have come about because students who had at least one experience of a test from the teacher before the summative examination were able to better judge what preparation would be most valuable for the summative examination. On average, effect sizes were smaller for longer treatments, probably because most longer

treatments had included at least one intermediate test for the less frequently tested group.

Bangert-Drowns et al. were surprised to find that the number of tests per week given to the high frequency group was not significantly correlated with the effect size. Rather, the effect size was best predicted from the frequency of testing for the control group. This suggests that the prime benefit from testing during a course comes from having at least one or two such tests, but that greater frequencies do not convey much benefit. One further analysis they conducted, however, raises some doubts about this conclusion. They identified eight studies which had groups with high, intermediate, and low frequencies of testing. Compared to the low frequency groups, the high frequency groups had a mean effect size of 0.48, whereas the intermediate frequency groups had a mean effect size of 0.22. The difference between 0.48 and 0.22 was statistically significant. This finding must be treated with caution, however, because of the small proportion of the total sample included in this analysis.

Overall, the evidence suggests that a moderate frequency of testing is desirable, and more frequent testing may produce further modest benefits. Groups that received no testing during the course were clearly disadvantaged, on average. Only four studies reported student attitudes towards instruction, but all favored more frequent testing, with a mean effect size of 0.59, a large effect.

One issue not covered in the review by Bangert-Drowns et al. was whether the tests during the course were cumulative, or related only to content since the last test (Keys, 1934; Rohm, Sparzo, & Bennett, 1986). The literature on distributed practice (see, for instance, Bjork, 1979, and Dempster, 1987) suggests that the use of cumulative tests, requiring repeated review of earlier material, would be advantageous on a comprehensive end of course examination. The extent of this benefit should depend on the nature of the course and of the examination. Hierarchical courses, in which later topics draw heavily on earlier material, tend to build in distributed practice, and thus should not benefit as much from directly retesting earlier content later in the course. On the other hand, courses that consist of a collection of topics that are only modestly interrelated would seem likely to benefit more from cumulative testing practices (see, for example, Guza & McLaughlin, 1987, who studied performance on spelling tests).

Another issue that needs further investigation is the effect of frequent testing on higher cognitive level outcomes. In their review, Bangert-Drowns et al. did not distinguish among studies by the cognitive level of the tests and criterion examinations, and it seems likely that most of the studies did not use significant numbers of questions at the higher cognitive levels of Bloom's taxonomy. It can be argued that frequent testing may not help (and may actually inhibit) higher level outcomes, even when the evaluations focus heavily on these outcomes. Students may need some "breathing space" if they are to adopt the deep level approaches that lead most effectively to higher level outcomes (Entwistle & Ramsden, 1983; Ramsden, 1985).

Effects of evaluative standards. The effects of teacher evaluative standards on student effort have been examined in a recent book by Natriello and Dornbusch (1984). They found that higher standards generally led to greater student effort and to students being more likely to attend class. Students who perceived standards as unattainable, however, were more likely to become disengaged from school. As

Natriello (1987) has suggested, there may well be a curvilinear relationship between the level of standards and student effort and performance, with some optimal level for each situation. This optimal level would probably depend on other aspects of the evaluation arrangements, such as whether or not students are given opportunities to get credit for correcting the deficiencies of evaluated work, or the nature of the feedback on their efforts. The weaker students, who are most at risk in high-demand classrooms, may need considerable practical support and encouragement if they are to avoid disillusionment.

Not surprisingly, Natriello and Dornbusch found that if students thought the evaluations of their work were not important or did not accurately reflect the level of their performance and effort, they were less likely to consider them worthy of effort. This conclusion is consistent with the results of research on student attributions of the reasons for success or failure in educational tasks (discussed later in this paper).

An important issue is whether the standards adopted are to be norm-referenced, criterion-referenced, or based on the effort and improvement of individual students (Natriello, 1987). This choice appears to differentially affect the motivation and learning of different categories of students. For instance, norm-referenced evaluation tends to undermine the learning and motivation of students who regularly score near the bottom of a class, while posing much less risk to the top students. No clear consensus emerges from the literature to date, but Natriello (1987) suggests that self-referenced standards may be optimal for most students. All students can improve their knowledge, skills, and attitudes and have this verified through evaluation, but only some can score above the class median on a measure.

When student performance on achievement tests is the criterion, research has generally shown that higher standards lead to higher performance (e.g., Rosswork, 1977), although again a curvilinear relationship may be predicted. Most of the relevant classroom-based research derives from studies of mastery learning, and these will be reviewed in a later section.

The Impact of Other Instructional Practices Involving Evaluation

Effects of adjunct questions in learning from text. In contrast to the research in the previous section, much of the research reviewed in this section has been conducted in laboratory settings. The findings of this research, however, converge with findings from research on the use of conventional tests in educational programs. Thus I believe that it is appropriate and valuable to include this extensive body of research in this review.

Adjunct questions are questions inserted before, during, or after a written passage that students are to study. Some studies have allowed the students to review earlier material after they encounter an adjunct question, whereas others have not permitted such looking back. The adjunct questions may be factual or *higher level* questions, although definitions of higher level vary markedly (Carrier & Fautsch-Partridge, 1981). Their effects have been studied by examining the pace and intensity of students' reading of portions of the passage, and by testing students in a variety of ways and at a variety of times on the content of the passage. These tests have looked at the students' grasp of the content or skill directly covered by the adjunct questions, their grasp of closely related material not directly addressed

by adjunct questions, and their ability to answer questions on material in the passage that is unrelated to any adjunct question. Finally, researchers have examined the effects on these outcomes of providing various forms of feedback on the students' performance on the adjunct questions, immediately or at some later stage.

This body of research has been reviewed recently by Hamaker (1986) and Hamilton (1985), and the earlier review of Anderson and Biddle (1975) has been widely cited. These reviews formed the starting point for my review of this area.

Factual adjunct questions. Hamaker (1986) used meta-analytic techniques to review 50 experiments on the effects of factual adjunct questions. She found that factual adjunct questions considerably improved the performance of students on subsequent test items testing the same facts. The average effect size was about 1.0 (the adjunct question group mean averaged one standard deviation higher than the control group mean). The mean effect size was very similar for adjunct questions placed before or after the relevant portion of the reading passage (prequestions or postquestions). The effect of these same factual questions on performance on test items covering related but not identical content was also positive, but of about half the magnitude (effect size approximately 0.5). Again, the mean effect size was similar for prequestions and postquestions. When Hamaker examined the effects of factual adjunct questions on unrelated test questions, she found modest negative effects (effect size approximately −0.3) for prequestions, and negligible effects for postquestions. The negative effect for factual prequestions has been interpreted by numerous researchers (e.g., Anderson & Biddle, 1975; Hamaker, 1986; Wittrock, 1986) as resulting from selective attention to the material cued by the prequestions. Similar effects have been found when students are given lists of factual objectives before a reading assignment (Hamilton, 1985; Wittrock, 1986). The negligible effect size on unrelated questions when factual postquestions are used differs from the findings of several previous reviews (e.g., Anderson & Biddle, 1975) that claimed a general facilitative effect of factual postquestions.

Hamaker reported that the effect sizes were unrelated to subjects' age, the interval between reading task and posttest, the average distance between adjunct questions and relevant text information, and whether or not subjects were allowed to consult the text while answering the adjunct questions. This final finding is important, even if based on relatively few studies, because most adjunct question studies have not allowed their subjects to look back to earlier portions of the text, a condition that reduces ecological validity for applying the conclusions about postquestions to normal studying (Duchastel, 1979a, 1983).

The format of the adjunct questions appears to have some influence on performance on test questions (Anderson & Biddle, 1975; Foos & Fisher, 1988; Hamaker, 1986). Adjunct questions in short-answer format have produced mean effect sizes about twice as large as adjunct questions in multiple-choice format, when performance on the same information was the criterion. Smaller benefits may also occur with related questions. This effect would seem to be due to the different processing demands of short-answer and multiple-choice adjunct questions.

The effect sizes were also found to be related to text length and the density of adjunct questions. The length of text and the ratio of length to number of questions were both positively correlated with effect size in studies using postquestions, but length was negatively correlated with performance on repeated test questions in studies using prequestions. It appears that the selective attention benefits of pre-

questions decrease as the number of facts to be covered and the amount of searching required increases, whereas postquestions work well with long texts, especially if the number of adjunct questions is not too large.

The beneficial effects of factual adjunct questions are not due to greater study time of students receiving adjunct questions. Although it is true that the inclusion of adjunct questions tends to increase study time a little where study time is not controlled, the effect sizes from studies in which study time was controlled (identical for experimental and control groups) were generally higher than the effect sizes from studies in which study time was not controlled (Hamaker, 1986, Table IX).

Higher order adjunct questions. Studying the effects of higher order adjunct questions is more complicated. These questions can be at a variety of cognitive levels and can require the student to integrate ideas from greater or smaller sections of the passage. The nature of the criterion questions is also important because greater effects would be expected with higher order criterion questions, but the effects on performance on factual criterion questions are also of interest (Watts & Anderson, 1971).

Hamaker included in her review 21 studies that compared the relative effectiveness of higher and lower order adjunct questions, and calculated her effect sizes based on this comparison. Compared to lower order questions, higher order questions led to substantially improved performance on test performance on the same higher order questions, and to moderately improved performance on related and unrelated higher order questions. They also improved performance a little on unrelated factual questions (Hamaker interprets this improvement as due to more thorough reading of the whole text). Compared to lower order questions, however, the higher order adjunct questions led to moderately lower performance on the content of the lower order questions, and on closely related content.

It should be emphasized that, as reported earlier, students receiving lower order adjunct questions generally performed significantly better than students who received no adjunct questions, so the comparative performance reported above does *not* mean that the use of higher order adjunct questions depresses performance on factual questions compared to a control group that receives no adjunct questions. Indeed, studies by Mayer (1975) and Watts and Anderson (1971), which included comparisons of groups answering higher order adjunct questions and control groups receiving no adjunct questions, found that the groups receiving higher order adjunct questions generally performed as well or better than the control groups on the factual test items. Thus it appears that the use of higher order adjunct questions is not detrimental to factual learning, but is distinctly advantageous to learning of higher order skills, whether directly covered by the adjunct questions or not. Hamaker concluded that higher order questions have a more general facilitative effect than factual questions. The results of Shavelson, Berliner, Ravitch, and Loeding (1974) suggest that this may be especially true for longer term retention.

D. W. Rowe (1986) reviewed evidence about the positioning of higher order questions, and concluded that the facilitative effects of higher order questions apply to prequestions as well as to postquestions. This conclusion may only apply, however, if students return to the questions and actively answer them.

Evaluation and the consolidation of learning. Beginning with early studies by Jones (1923) and Spitzer (1939), numerous studies have demonstrated that taking a test on a topic after studying it tends to enhance longer term retention of the

material studied, even if no feedback is given on the test performance. In many cases, the observed effect has been strong (for instance, Jones, 1923, found retention test scores for tested students double those for untested students). This effect has been described as the consolidation function of testing, and would appear to closely parallel the benefits of adjunct postquestions (Duchastel, 1979b, Duchastel & Nungester, 1982, Foos & Fisher, 1988). Indeed, where the learning is from reading, the effect is indistinguishable from the effect of adjunct postquestions in the special case where the postquestions are massed together at the end of the reading passage and students are not permitted to look back at the material they have read. Because Hamaker (1986) found no important difference between the effects of inserted and massed postquestions, the findings reported earlier for adjunct postquestions should also apply in this situation.

The benefits from testing can apparently be explained by three factors. First, the testing gets the students to attend to the content another time. This constitutes a limited form of distributed practice, and the beneficial effects of distributed practice on retention are well established. Second, the testing encourages the student to actively process content, which is known to enhance learning and retention (Brown, Bransford, Ferrara, & Campione, 1983; Levin, 1982; McKeachie, Pintrich, Lin, & Smith, 1986; Thomas & Rohwer, 1986; Wittrock, 1974, 1979, 1986). Some types of items may stimulate more active processing than others (Duchastel, 1981). Third, the test directs attention to the topics, skills, and details tested, which may focus the student's preparation for a subsequent retention test. Students are more likely to achieve goals that they clearly perceive (Anderson & Armbruster, 1984; Brown et al., 1983; Rohwer & Thomas, 1987; Thomas & Rohwer, 1986). All of these effects are predominantly associated with the content actually tested, so it is not surprising that little benefit has been shown for untested material unless it is closely related to the tested material (see, for instance, LaPorte & Voss, 1975; Nungester & Duchastel, 1982).

Effects of teacher oral questions in class. The extensive literature on the relationships between teacher classroom behaviors and student achievement (from process-product and experimental studies) has been summarized recently by Brophy and Good (1986) and Rosenshine and Stevens (1986). One aspect of teacher behavior they discuss is the use of teacher oral questions directed to students and the feedback given to student answers. Research on teacher questioning (often called recitation) has also been reviewed by Gall (1984).

These reviews report that the frequency of teacher questioning has generally been shown to be positively related to student achievement. Rosenshine and Stevens (1986) state that "the critical variable seems to be a high percentage of student responses" (p. 383). Reasons for the effectiveness of recitation have been suggested by Gall (1984). These include several factors already discussed in this review: that questions encourage more active engagement in learning; that they provide practice on the material, which helps to consolidate student learning; that they lead to feedback that clarifies understanding and corrects misconceptions; that they cue students as to the aspects the teacher regards as more important (and thus more likely to be included in tests subsequently); and that they give practice on activities similar to those in the criterion tests.

In order to obtain full benefit from classroom questioning, the reviewers suggest that questions should be directed to as many students as possible (to encourage all

toward active learning), that teachers need to practice phrasing questions in ways that communicate the task clearly, that the difficulty level should be such that the majority of questions receive satisfactory responses, and that responses to other than simple factual questions tend to be fuller and more appropriate if several seconds are allowed between question and response (see also M. B. Rowe, 1986). Feedback should include knowledge of results, but should make only limited use of praise (e.g., praise might be used mainly for correct responses from anxious or less capable students) and very little use of criticism.

Perhaps the most frequently researched aspect of teacher oral questions has been the cognitive level of the questions and the effects of different cognitive levels on student achievement. This has also been an area in which reviewers have reached markedly varied conclusions, although the reviewers have agreed that higher level questions are generally used much less than lower level questions (a ratio of 1 to 3 is typical of reported figures from research in school classrooms). Medley (1979) and Rosenshine (1979) both concluded that greater use of higher level questions led to lower student achievement. Winne (1979), in a review of relevant experimental studies, found no clear pattern of achievement change associated with greater use of higher level questions. Redfield and Rousseau (1981), however, used meta-analysis on a very similar collection of experimental studies and reported a mean effect size of 0.73 favoring use of higher cognitive level questions. More recently, Samson, Strykowski, Weinstein, and Walberg (1987) conducted another meta-analysis of experimental studies and found a mean effect size of 0.26 favoring use of higher level questions.

Several factors help to make sense of these contradictory findings (cf. Gall, 1984; Samson et al., 1987). First, studies in this area have been very inconsistent in their definitions of higher and lower level questions. Lower level has been defined to include the bottom one, two, or three categories from Bloom's taxonomy, and other taxonomies have also been used. Second, the difficulty of the questions has rarely been controlled, so that higher level questions may have been substantially more difficult on average than lower level questions, which could have reduced students' opportunity and motivation to learn effectively from these questions. Third, too little attention has been paid to the nature of the criterion achievement measures. The use of a criterion involving only factual recall or recognition could be predicted not to favor the use of higher level oral questions. For example, the studies reviewed by Medley (1979) and Rosenshine (1979) were predominantly conducted in junior elementary school classes with high proportions of disadvantaged children, where the teaching focused very much on basic knowledge and skills. These students may have had difficulty attending to and correctly interpreting the higher level questions, and the criterion measures used generally included few higher level questions. Fourth, many of the studies were of very brief duration. It could be predicted that higher cognitive level questions would be most effective when used consistently over substantial periods of time, especially if students had previously had little experience with such questions. This prediction is supported by an analysis included in the review by Samson et al. (1987). They found a mean effect size of 0.05 for 22 studies lasting 5 days or less, but a mean effect size of 0.83 for 4 studies lasting 20 days or more. Finally, it is interesting to note that the review by Samson et al. also reported markedly larger mean effect sizes in studies that were better designed (random assignment to treatments, sample size greater than

50), and in studies that more closely specified and/or monitored the degree to which higher level questions were used.

Taking all these considerations into account, I believe it is justifiable to conclude that the use of higher level oral questions by teachers usually fosters, or at least does not harm, student achievement. The main exceptions are likely to be situations in which the achievement measure consists almost entirely of factual recall or recognition questions and situations in which the higher level questions are too difficult or too unclear for many of the students. Careful guidance and training may be needed before some students can respond appropriately to higher level questions (see Dillon, 1982; Klinzing, Klinzing-Eurich, & Tisher, 1985; Mills, Rice, Berliner, & Rousseau, 1980). Further, if higher level questions are to substantially enhance student achievement, they will need to be used consistently over extended periods of time.

The impact of the cognitive level of questions on student affect has not received much attention. In particular, it seems reasonable to hypothesize that higher level questions of appropriate difficulty would tend to enhance student interest in the course content more than factual questions. Because of the long-term importance of motivational factors in learning, research is needed to investigate this hypothesis.

Effects of feedback on performance. There is extensive literature on the effects on providing knowledge of results and other forms of feedback on the evaluative tasks performed by students. Factors involved include the nature and extent of the feedback, its timing, its value in relation to the student's existing level of performance, and its relationship to the summative functions of evaluation.

Research that examined the effects of feedback on learning from text was reviewed thoroughly by Kulhavy (1977). He found that feedback generally increased what students learned from reading assignments that included questions or tests for them to answer.

One exception to this positive conclusion occurred if the material was too difficult for the students to process, so that they tended to choose to try to learn the highlights from the feedback. This exception is further supported by a recent meta-analysis of the effects of feedback in 22 studies involving programmed and computer-based instruction. In this meta-analysis, Bangert-Drowns, Kulik, and Kulik (1987) found a correlation of −0.44 between task difficulty (control group error rate) and benefit of feedback (effect size comparing feedback group mean with control group mean). Where error rates are high, the task of learning from the feedback apparently becomes daunting.

A second exception occurred if the feedback was available too soon (as in some programmed textbooks), thus allowing the student to avoid careful reading and answering of the questions. This exception has also been confirmed by Bangert-Drowns et al. (1987), who found in their sample that where students had to make responses to questions before receiving feedback, the average effect size was 0.38, but where feedback was available without a student response, the average effect size was −0.13.

Research on feedback on learning from classroom teaching has produced similar findings on the effectiveness of feedback (Beeson, 1973; Bergland, 1969; Ingenkamp, 1986; Karraker, 1967; O'Neill, Razor, & Bartz, 1976; Page, 1958; Sassenrath & Garverick, 1965; Strang & Rust, 1973; Wexley & Thornton, 1972).

Functions and form of feedback. Kulhavy reported that feedback acts to confirm

correct answers, thus helping students to "know what they know." There is very little evidence that such knowledge of correct responses acts by reinforcing the correct response, and indeed feedback on correct responses has little effect on subsequent performance, except perhaps in the special case where the student has grave doubts about the correctness of the initial answer.

The major benefit from feedback reported by Kulhavy is the identification of errors of knowledge and understanding, and assistance with correcting those errors. In most studies, such feedback clearly improved subsequent performance on similar questions. Feedback on incorrect responses has been shown to be most effective where the initial response was made with high confidence, probably because the student attends more to the feedback in such cases (due to the element of surprise and the initial desire to defend the correctness of the response).

It seems likely that the most effective form of feedback will depend on the correctness of the answer, the student's degree of confidence in the answer, and the nature of the task. If the answer is correct, simple confirmation of its correctness is sufficient. If the question was factual and the answer is incorrect, the most efficient form of feedback is probably simply to give the correct answer (Phye, 1979). If the question involves comprehension or higher cognitive skills, however, more detailed feedback is desirable. Students who answered such questions incorrectly with high confidence may need help to identify the source of their misunderstanding (Block & Anderson, 1975; Fredericksen, 1984b), whereas students who answered the question incorrectly with low confidence may need to be given conceptual help and advised to restudy the material.

There is little support from laboratory or classroom research for making praise a prominent part of feedback, but Page (1958) found that simple positive comments were beneficial, and harsh criticism is predictably counterproductive. Both the age and achievement level of the student may modify this conclusion: younger and less able students may benefit most from praise. Praise should be reserved for specific achievements that truly represent substantial accomplishments for the individual student. The motivational effects of different types of feedback are discussed in more detail in later sections of this paper.

Feedback can also play a very positive role in guiding students in their use of learning strategies (Pressley, Levin, & Ghatala, 1984). Pressley et al. found that explicit feedback on strategy use was especially valuable with young children, whereas adults who had tried several strategies and been tested on their learning were generally able to identify the most effective strategy.

The timing of feedback. Effects of the timing of feedback have received considerable attention. Kulik and Kulik (1988) used meta-analytic techniques to review 53 studies of the timing of feedback in verbal learning. They identified three different categories of study, finding quite different results for the three categories. A key factor that apparently influenced these differences was whether or not the criterion test questions were identical to the earlier feedback questions. Where different questions were used, most studies found a small advantage for immediate feedback (the mean effect size for 11 studies was 0.28). Where identical questions were used (e.g., Kulhavy & Anderson, 1972), however, most studies found a modest advantage for delayed feedback (the mean effect size for 14 studies was −0.36). Kulhavy and Anderson (1972) suggested that this effect arose because the memory of incorrect responses made during acquisition interfered with the learning of the

correct responses from the immediate feedback. Such interference could be expected to decrease with delayed feedback, which would essentially serve as a second learning trial, providing distributed practice on the task.

In most classroom situations, where the tasks leading to feedback form only a sample of the desired course outcomes, these data suggest that immediate feedback will be more beneficial than delayed feedback. Because the typical effect sizes were not large, however, the precise timing of feedback does not appear to be too critical, unless it is delayed so long that students have little motivation to pay close attention and learn from it.

Are feedback and summative evaluation compatible? A final issue to be addressed here is whether the feedback and summative purposes of student evaluation are best separated. Strong arguments for such separation have been presented by McPartland (1987), Miller (1976), Sadler (1983), and Slavin (1978), among others. They argue that where evaluations count significantly toward the student's final grade, the student tends to pay less attention to the feedback, and thus to learn less from it. This effect should be reduced if students are given multiple opportunities to test and prove their achievement, with only the final evaluation counting toward their grade, as is generally the case in mastery learning procedures. Of course, one argument for counting more evaluations in grading is to improve the reliability of the grading process, but this consideration will often be less important than the benefits of evaluation for learning.

Effects of mastery testing. Kulik and Kulik (1987) conducted a meta-analysis of studies of testing in mastery learning programs, analyzing data from 49 studies. Each study took place in real classrooms, provided results for both a class taught with a mastery testing requirement and a class taught without such a requirement, and was judged free of serious experimental bias. The studies varied in length from 1 to 32 weeks, with about half shorter than 10 weeks. Effect sizes were again used to describe the findings.

The mean effect size on summative, end-of-course examination performance was 0.54, a strong effect. Kulik and Kulik note that this mean effect size is substantially lower than the figure of 0.82 reported recently by Guskey and Gates (1986) in another review of studies on mastery learning. They rightly point out, however, that 9 of the 25 studies used by Guskey and Gates calculated effect sizes using combined scores from the instructional quizzes (which the mastery groups had multiple opportunities to pass) and the final examination, thus biasing the results in favor of the mastery group. The mean effect size for the 16 studies which avoided this bias was 0.47, a figure much more consistent with the Kuliks' findings.

Effect sizes varied markedly in relation to three features of the studies. Studies that had the same frequency of testing in both groups had a mean effect size of 0.48, whereas studies in which the test frequency was not controlled (usually higher in the mastery testing group) had a mean effect size of 0.65. This difference was not statistically significant, but it is worthy of note that the extra benefit for more frequent testing is similar to the 0.25 reported in the earlier section on frequency of testing.

A statistically significant difference was found between effect sizes from studies in which similar levels and types of feedback were given to students in both groups, and those from studies in which this was not the case (in these cases, the mastery testing groups could be expected to have received more feedback). The two mean

effect sizes were 0.36 and 0.67, suggesting that a major component of the effectiveness of mastery testing arises from the additional feedback that it usually provides.

The other statistically significant difference was between studies at varying levels of mastery criterion. In 17 studies where the criterion level for mastery was a score of 91% or higher on unit tests, the mean effect size was 0.73; in 15 studies with a criterion level of 81 to 90%, the mean effect size was 0.51; and in 17 studies with a criterion level below 81% the mean effect size was 0.38. This is a strong effect, demonstrating that under mastery testing conditions a higher criterion level generally produces greater learning (assessed on an end-of-course examination).

Thus the results of research on mastery testing suggest that the sizeable benefits observed largely represent the combined effects of the benefits described in earlier sections from more frequent testing, from giving detailed feedback on their progress on a regular basis, and from setting high but attainable standards. One further effect that is probably important is the benefit of allowing repeated opportunities to attain the standards set. This feature might have considerable benefits in increasing motivation and a sense of self-efficacy, while reducing the anxiety often associated with one-shot testing (Friedman, 1987). Kulik and Kulik (1987) reach a similar conclusion to Abbott and Falstrom (1977): the other features often included in courses based on mastery learning models do not appear to add significantly to the effects described above.

As in the section on frequency of testing, some caution must be expressed about the generalizability of the findings on mastery testing because the cognitive levels of the tests and examinations were not analyzed. Different effects may occur for courses and tests that heavily emphasize higher cognitive level outcomes, especially in relation to the benefits of more frequent testing. The benefits of feedback, of opportunities for extra attempts at tasks initially handled poorly, and of challenging standards seem more likely to apply to evaluation tasks at all cognitive levels.

Effects of competitive, individualistic, and cooperative learning structures. Many studies have examined the effects of different classroom learning and goal structures on students. In particular, considerable attention has been given to the effects and comparative merits of competitive, individualistic, and cooperative learning structures. In competitive structures, the success or failure of students is largely determined by their performance relative to other students. In individualistic structures, students are rewarded on the basis of their own work, independent of the work of other students. In cooperative structures, students work together in groups, and judgments of success are based on the overall achievements of each group. Ames (1984) has classified these situations according to the pattern of interdependence among students. Competitive structures involve negative interdependence because success for one student reduces the chances that other students will succeed. In individualistic structures, there is no interdependence among students. Finally, in cooperative structures, there is positive interdependence among students, since success for one student assists the success of all members of the group in which that student is a member.

Effects on cognitive outcomes. Johnson, Maruyama, Johnson, Nelson, and Skon (1981) conducted a meta-analysis of 122 studies that examined the comparative effects on student achievement of two or more of these categories (for their purposes, they identified four categories, subdividing the cooperative structure category

into two subcategories: cooperation with intergroup competition, and cooperation without intergroup competition). They used three different ways of summarizing the findings (vote count, a z-score method, and effect size). Because these three approaches usually produced similar conclusions, I shall base my summary of their findings on the effect size data. They found that competitive and individualistic structures seemed equally effective, with a mean effect size between these structures of 0.03. Cooperative structures (without intergroup competition) generally produced higher achievement than competitive or individualistic structures (both mean effect sizes were 0.78). Structures that involved cooperation within groups but competition between groups also led to higher average achievement than competitive or individualistic structures (mean effect sizes of 0.37 and 0.50, respectively). Johnson et al. (1981) also conducted regression analyses to examine the influence on these effect sizes of some 20 possible mediating or moderating variables, although small sample sizes restricted the usefulness of many of the findings. There was some evidence that the benefits of cooperative structures were greater when group sizes were small (2 or 3), when the task required more interdependence among group members (e.g., a group product was to be generated), and when the task was not a simple exercise (see Johnson, Maruyama, & Johnson, 1982, for the clearest data on these issues). Overall, Johnson et al. (1981) concluded that cooperative structures are generally superior to competitive or individualistic structures in promoting student achievement.

This conclusion was criticized by Cotton and Cook (1982) and McGlynn (1982), with a response from Johnson et al. (1982). The heart of the criticism involved concern that no such general statement could be made, given the reported interactions of the effect sizes with other variables, and further probable interactions with other variables that were not studied. Johnson et al. (1982) effectively refuted some of the more specific criticisms, but agreed that there probably are learning situations in which cooperative structures are not as effective as competitive or individualistic structures. They noted, however, that such situations appear to be much less common than those in which cooperative structures are superior.

The effects on achievement of cooperative learning structures have been further analyzed by Slavin (1983b, 1984), who focused on the value of cooperative incentives. Cooperative incentives are incentives in which the rewards for individuals are based on performance of the group as a whole (either through a group product or through the aggregated performances of the individual group members). Slavin contrasted three incentive situations for students who have been asked to work on tasks in groups: group reward for the individual performance of group members, group reward for a group product, and individual reward for performance tested individually after the group activities were completed. He reported strong evidence (based on 28 studies) that the use of group reward based on the individual performance of group members was an effective strategy for enhancing the mean achievement of the group, hypothesizing that this incentive structure encouraged group members to be concerned about improving the learning of all group members. Slavin reported that studies of the use of group reward on the basis of a group product did not demonstrate any clear superiority of cooperative learning over noncooperative approaches. Slavin gave some emphasis to this finding in his conclusions, causing some controversy because it was based on only eight studies

and should thus probably be regarded as tentative. The 10 studies in which individual rewards were given based on individual performance showed no advantage for cooperative study over noncooperative approaches.

Slavin concluded that the use of group rewards based on the individual performance of group members is essential to the effectiveness of cooperative learning methods. Such a strong conclusion may not be justified on the basis of the data he reported, but this incentive structure does appear to be beneficial to group learning (see also Lew, Mesch, Johnson, & Johnson, 1986).

Effects on social outcomes. One widely cited benefit of cooperative learning structures is that they lead to increased cohesiveness among the students involved (Johnson, Johnson, & Maruyama, 1983; Slavin, 1983a). This can be especially beneficial in classes that are diverse in ethnic composition, ability level, or because of the inclusion of mainstreamed handicapped students. Johnson et al. (1983) conducted a meta-analysis of 98 studies of cooperative learning, with interpersonal attraction as the dependent variable. They found little difference between competitive and individualistic structures, but students in cooperative structures scored substantially higher in mean interpersonal attraction. Where the cooperative groups were not competitive with each other, the effect size was 1.11 (compared both to competitive and to individualistic structures). Where there was competition between groups, the mean effect size was smaller (0.79 compared to individualistic structures, 0.55 compared to competitive structures). Clearly, structures that encourage cooperation among students can have substantial beneficial effects on social relationships among students.

Astin (1987) discussed the benefits of cooperative learning in higher education. Among other things, he emphasized that a key benefit could be an enhanced sense of mutual trust, both among students and between students and teacher. He noted that in competitive learning situations, students often work very hard to disguise their ignorance (from peers and from their teacher). This limits the availability and effectiveness of feedback, thus undermining learning. Astin sees cooperative structures helping to overcome this problem, while fostering interpersonal skills that are greatly needed in the community.

Motivational Aspects Relating to Classroom Evaluation

Research has repeatedly demonstrated that the responses of individual students to educational experiences and tasks are complex functions of their abilities and personalities, their past educational experiences, their current attitudes, self-perceptions and motivational states, together with the nature of the current experiences and tasks. Effective education requires the fusing of "skill and will" (Paris, 1988; Paris & Cross, 1983), and intrinsic interest and continuing motivation to learn are educational outcomes that should be regarded as at least as important as cognitive outcomes (Maehr, 1976; Paris, 1988). The importance of motivational factors has been vigorously stated by Howe (1987):

> I have a strong feeling that motivational factors are crucial whenever a person achieves anything of significance as a result of learning and thought, and I cannot think of exceptions to this statement. That is not to claim that a high level of motivation can ever be a sufficient condition for human achievements, but it is undoubtedly a necessary one. And, conversely, negative motivational influences, such as fear of failure, feelings of helplessness, lack of confidence, and having the

> experience that one's fate is largely controlled by external factors rather than by oneself, almost certainly have effects that restrict a person's learned achievements. (p. 142)

Modern theories of achievement motivation (Dweck & Elliott, 1983; Eccles, 1983; Nicholls, 1984; Weiner, 1986) place considerable stress on the importance of student self-perceptions in determining responses to educational and evaluative tasks. Thus, for instance, the attributions (reasons) students give for their success or failure, or their perceptions of self-efficacy (capability to perform well) are highly important factors influencing their behavior. To a significant degree these variables are task- or domain-specific, so that it is more profitable to think about them in this way than as enduring general characteristics.

One important factor that should be taken into account in considering the relationship between motivational variables and achievement is the repeated finding or suggestion of curvilinear relationships (Eccles, 1983; McKeachie et al., 1986). Both very high and very low levels on motivational variables may be less desirable than intermediate levels. For instance, if perceived task importance is very low, many students may not try very hard (as reported by Natriello & Dornbusch, 1984). If perceived task importance is very high, however, anxiety may inhibit performance (Tobias, 1985). Similarly, if students have a very low level of self-efficacy for a task, they are unlikely to attack the task with much enthusiasm or persistence, but if they have a very high level of self-efficacy, they may not give the task sufficient care and attention to achieve good results (Schunk, 1984).

The following subsections briefly review five interrelated areas of research on student motivation and affect. In each area, the classroom evaluation of students appears to play a major role. Although motivational considerations are emphasized, effects on cognitive outcomes are also discussed where appropriate.

Test anxiety. The research on test anxiety has been reviewed by Hill (1984), Hill and Wigfield (1984), McKeachie (1984), McKeachie et al. (1986), Sarason (1980), and Tobias (1985). Studies have repeatedly shown substantial negative correlations between measures of test anxiety collected before tests are administered and performance on those tests. The magnitudes of the correlations appear to increase at higher grade levels, with one study finding a correlation as strong as $-.60$ for 11th-grade students (see Hill, 1984, p. 248). The debilitating effects for high anxiety students are greater when the student perceives good performance on the test to be particularly important, when the test is expected to be difficult, and when the testing conditions are particularly intrusive (e.g., rigid time limits and associated time pressures, special test instructions or conditions, unfamiliar test formats). Thus the effects tend to be greater on standardized tests than classroom tests.

Although failures on earlier tasks clearly influence the development of anxiety, the anxiety does not simply arise from lack of the knowledge or skills required to answer the test items. Several studies have shown that high anxiety students do much better on the same cognitive tasks administered under less stressful conditions, performing at levels much closer to those of their less anxious peers (Hill, 1984; Hill & Wigfield, 1984).

A number of different mechanisms have been suggested to explain the debilitating effects of anxiety on achievement (McKeachie et al., 1986; Tobias, 1985). These have included suggestions that high anxiety students may be weak in their use of

cognitive and metacognitive learning strategies, that they may use poor test-taking strategies, or that they may be particularly prone to distracting thoughts while taking a test (such as thoughts about failure or about difficult items yet to be completed). These proposed mechanisms are clearly not mutually exclusive. The first (weak learning strategies) would not explain the findings reported in the last paragraph, so it is not a sufficient explanation by itself. However, it does have empirical support (Naveh-Benjamin, McKeachie, & Lin, 1987). The other two mechanisms are more specific to the testing situation, and both have empirical support.

Several guidelines have been suggested for reducing the debilitating effects of test anxiety in classroom evaluation programs (Hill, 1984; Hill & Wigfield, 1984). These include: testing under "power" testing conditions (very generous time limits, so no student feels under significant time pressure); avoiding distinctive and stressful testing conditions; giving the students ample details of the nature, difficulty, and format of the test (with practice examples); setting tasks that allow each student a reasonable level of success; reducing emphasis on social comparison (Hill & Wigfield suggest avoiding the use of letter grades in elementary schools); and providing special training for students who may be victims of test anxiety.

Student self-efficacy. Self-efficacy, as defined by Bandura (1977, 1982), refers to students' perceptions of their capability to perform certain tasks or domains of tasks. Research on the role of self-efficacy in achievement behavior and classroom learning has been reviewed by Schunk (1984, 1985). Perceptions of self-efficacy in an area have been shown to correlate highly with achievement in that area. For instance, in a recent study by Thomas, Iventosch, and Rohwer (1987), self-efficacy was found to be a better predictor of school achievement than their selected measure of academic ability. They also found that students with high self-efficacy tended to make more use of deeper learning strategies (generative and selective activities) than other students did.

Perceptions of self-efficacy appear to have a strong influence on effort and persistence with difficult tasks, or after experiences of failure (Bandura, 1982; Schunk, 1984, 1985). Under such circumstances, students high in self-efficacy usually redouble their efforts, whereas students low in self-efficacy tend to make minimal efforts or avoid such tasks.

The main mechanism for building self-efficacy in a particular domain appears to be experiencing repeated success on tasks in that domain. Success at tasks perceived as difficult or challenging is more influential than success on easier tasks. On the other hand, of course, repeated failure leads to lowered self-efficacy. More than 40 years ago, E. L. Thorndike began a paper with these words:

> It is a matter of common knowledge that a mind which for any reason becomes engaged in an activity and finds itself repeatedly and persistently failing therein, is impelled to intermit or abandon it. The person does abandon it unless this impulsion is counterbalanced by some contrary force, such as the hope of a turn of the tide toward success, or an inner sense of worth from maintaining the activity, or a fear that worse will befall him if he stops. (Thorndike & Woodyard, 1934, p. 241).

To foster self-efficacy, evaluations of task performance should emphasize performance (task mastery) rather than task engagement (Schunk, 1984). Thus, for

instance, grade credit should be given for quality of work on an assignment, not merely for handing it in (Schunk, 1984). Also, the emphasis in performance feedback should be on informing students about their progress in mastery, rather than on social comparison (Schunk, 1985). This is crucial for the less able students, who might otherwise receive little positive feedback. Finally, there is strong evidence that self-efficacy is best enhanced if longer term goals are supported by a carefully sequenced series of subgoals with clear criteria that students find attainable. This is especially important if the students are young, or if they initially lack confidence or interest in the domain (Bandura & Schunk, 1981). These requirements are met by mastery learning procedures, if well implemented (Driscoll, 1986), but can also be incorporated in other approaches to teaching and learning. One concern is that teaching and evaluation arrangements be sufficiently flexible to ensure suitably challenging tasks for the most capable students, as otherwise they would have little opportunity to build their perceptions of self-efficacy (and much opportunity for boredom).

Intrinsic motivation and continuing motivation. Intrinsic motivation to learn (defined as a self-sustaining desire to learn) and continuing motivation (defined by Maehr, 1976, as a tendency to return to and continue working on tasks away from the instructional context in which they were initially confronted) are highly related concepts. Both, in turn, are closely related to interest in the material that is being studied. Few would disagree that such interest is a very desirable outcome of educational activity, and also a very important factor influencing the quality and extent of learning activities. Alfred North Whitehead, in his characteristically forthright way, went so far as to suggest that "there can be no mental development without interest. Interest is the *sine qua non* for attention and apprehension" (Whitehead, 1929, p. 48). Maehr (1976) argues that continuing motivation is also important because learning does not just take place in classrooms. Activities that students engage in by choice outside the classroom can complement and strengthen classroom-based learning, and can also lead to that learning being extended and updated long after the formal classroom program ends.

The research on intrinsic and continuing motivation has been reviewed in recent years by Corno and Mandinach (1983), Corno and Rohrkemper (1985), deCharms (1976), Deci (1975), Deci and Ryan (1985), Harter (1985), Maehr (1976), McCombs (1984), and Ryan, Connell, and Deci (1985), among others. Corno and her colleagues have argued that intrinsic motivation and self-regulated learning are closely linked, presenting evidence that self-regulated learning experiences foster intrinsic motivation, and that intrinsic motivation in turn encourages students to be more independent as learners. There is general agreement among the other reviewers that allowing a degree of student autonomy in choice of learning activities and objectives is a key factor in fostering intrinsic motivation. In considering the problem of passive reading failure, Johnston and Winograd (1985) drew on the work of deCharms (1983) and others to suggest that more opportunity might be given in school for students to engage in recreational reading. They point out that this would both encourage and make use of intrinsic motivation, and that it would also reduce the likelihood of normative social comparisons, because any evaluation of this activity would of necessity have to be individualized.

There is also widespread agreement among the reviewers that the use of extrinsic motivation is problematic. The problems can be illustrated by briefly examining

the findings of three studies. Lepper, Greene, and Nisbett (1973) found that students who had previously chosen to engage in an activity voluntarily, with apparent enjoyment, were less inclined to return to that activity after they had received a reward from a teacher for engaging in the activity. Maehr and Stallings (1972) studied students performing easy or hard tasks under extrinsic or intrinsic motivation conditions. They found that students who worked under the intrinsic motivation condition continued to be interested in working on difficult tasks, whereas students who worked under the extrinsic motivation condition lost interest in attempting difficult tasks, preferring to attempt only easy ones (see also Hughes, Sullivan, & Mosley, 1985). Finally, Condry and Chambers (1978) found that students in their extrinsic motivation group were more *answer oriented*, trying to take shortcuts to produce the desired answers, whereas students in the intrinsic motivation group tended to use deeper, more meaningful approaches to understanding the tasks.

These and other studies have repeatedly shown that where students are initially intrinsically motivated, attempting to stimulate learning through extrinsic motivation usually leads to decreased intrinsic motivation, especially on challenging tasks. Such a result is clearly not desirable. On the other hand, where students initially lack intrinsic motivation in a particular subject area, research reported in the last section suggests that a carefully planned program of positive educational experiences accompanied by extrinsic motivation can lead to the development of interest in the area, and thus to intrinsic motivation. Unfortunately, however, there is strong evidence that in most education systems such gains are usually outweighed by the losses. Many observers have commented on the contrast between the broad enthusiasm for learning demonstrated by most children in the first year or two of schooling and the jaded approach of many older students. Although some of this difference may relate to developmental factors, it is hard to escape the conclusion that for many students schooling tends to lower rather than increase interest in learning.

It is important to note that classroom evaluation procedures need not have the debilitating effects on intrinsic motivation noted above. Deci (1975) and others (Keller, 1983; Ryan, Connell, & Deci, 1985) have noted that the key factor seems to be whether students perceive the primary goal of the evaluation to be controlling their behavior or providing informative and helpful feedback on their progress in learning. Evaluation can be used as a bludgeon to make students learn, and in the short term this may produce significant learning, but the longer term consequences of such an approach appear to be most undesirable, especially for the less able students.

Attributions for success and failure. Extensive research has demonstrated that student self-perceptions of the factors influencing success or failure in learning tasks have a very significant influence on their motivation and behavior. Such attributions for success or failure are central to Weiner's theory of achievement motivation (Weiner, 1979, 1985, 1986), and many other researchers on motivation have also stressed their importance. Research on student attributions has been reviewed by Covington (1984, 1985), Dweck and Elliott (1983), Nicholls (1983, 1984), Paris and Cross (1983), and Weiner (1985, 1986), among others.

Weiner (1979) stated that success or failure could be attributed to four possible causes: ability, effort, luck, or task difficulty. The first two of these are internal to

the student, the latter two are external. Weiner also identified emotional consequences when success or failure is attributed to these causes. For instance, he stated that success which is attributed to ability or effort leads to pride and self-esteem, that failure attributed to lack of effort leads to guilt, and that failure attributed to stable factors (lack of ability or task difficulty that is consistently too high) leads to hopelessness.

Nicholls (1984) and Dweck and Bempechat (1983) reviewed evidence that students do not share a single conception of ability. Up to about 10 years of age, students generally conceive of ability as learning through effort, so that gains in task mastery are indicative of enhanced ability. Many older children and adults, however, conceive of ability as a stable trait that is judged normatively (i.e., by comparing performances of different individuals). In this conception, normatively superior performance is indicative of ability, especially if it requires comparatively little effort.

These two conceptions of ability differentially affect the achievement behavior of students (Covington, 1984, 1985; Nicholls, 1984). Students with the task mastery concept of ability like challenging tasks that appear reasonably likely to yield success after considerable effort. Such tasks can give them a sense of achievement and thus enhance their perceived ability. Among students with the normative concept of ability, those who believe they are of high ability tend to prefer tasks that they perceive as of medium difficulty, and thus are likely to confirm their ability by again distinguishing them from students of lower ability. On the other hand, those who believe they are of low ability try to avoid tasks of medium difficulty, because these are likely to confirm their low ability by requiring substantial effort yet carrying a substantial risk of failure. Such students prefer either easy or difficult tasks because these are less likely to demonstrate their lack of ability. Students who perceive themselves as having insufficient ability to do well on most assigned classroom tasks tend to display helplessness (see Dweck & Elliott, 1983). This means that they do not expend much effort to learn because they expect their efforts to result in failure anyway, and thus to reemphasize their low ability.

Several researchers (Ames, 1984; deCharms, 1983; Maehr, 1983; Nicholls, 1983) have identified two or more categories of achievement goals. They all make a distinction between task goals and ego goals, which parallel the two conceptions of ability discussed above. With task goals, students believe they are responsible for the outcome, that there are reasonably clear mastery criteria for success, and that the outcome is not preordained. Task goals are often associated with intrinsic motivation. With ego goals, on the other hand, the key feature is that success requires doing better than someone else.

Task characteristics interact with students' personal conceptions of ability to determine whether students treat particular tasks as task goals or ego goals. For instance, many computer games are extremely challenging for the players, yet less competent players often display high levels of motivation and persistence because the task characteristics favor task goals and reduce the salience of ego goals. By contrast, many other tasks give much less criterion-referenced feedback, and under such conditions success is more likely to be judged normatively.

This research seems to have clear implications for classroom teaching and evaluation. If all students are to be encouraged to learn, conditions that favor task goals over ego goals are desirable. These conditions include challenging but attain-

able tasks, some individualization of tasks, use of tasks that are more intrinsically motivating or more gamelike in nature, opportunities for student autonomy in learning, little use of ability groups, use of cooperative learning approaches, provision of unambiguous performance feedback that emphasizes mastery and progress (rather than normative comparisons), and little emphasis on summative grading (Covington, 1985; Johnston & Winograd, 1985; Maehr, 1983; Nicholls, 1983; Rosenholtz & Simpson, 1984). Under such conditions, failure at a task is more likely to be constructive rather than destructive (Clifford, 1984). If such conditions could be fostered, perceived ability stratification would be reduced, with consequent reductions in the serious differential changes of self-esteem that occur from about the age of 10 (Kifer, 1977).

Motivational aspects of competitive, individualistic, and cooperative learning structures. Research on motivational aspects of competitive, individualistic, and cooperative task and incentive structures has been reviewed by Ames (1984), Johnson and Johnson (1985), and Slavin (1987). The motivational effects of competitive structures have been discussed in earlier sections, but will be briefly summarized here. Social comparison (norm referencing) is central to competitive structures. This tends to result in severe discouragement for the students who have few academic successes in competition with their peers. It discourages students from helping each other with their academic work, and also threatens peer relationships, encouraging an "us and them" mentality which tends to segregate the higher and lower achieving students (Deutsch, 1979). It does not encourage intrinsic motivation. Finally, it tends to encourage students to attribute success and failure to ability rather than to effort, which is especially harmful for the weaker students.

In individualistic structures, rewards are based on criterion-referenced evaluation. If all students are evaluated on the same tasks, using the same standards, this can simply become another type of competitive structure (Ames, 1984), but at least there is some possibility of all students meeting specified passing standards. The provision of repeated opportunities to meet the standards can be a key factor in reducing the competitiveness of such individualistic structures. If, on the other hand, student's programs of work are more individualized, and the emphasis in evaluation is placed on each student's progress in learning, competitiveness is minimized. Under these circumstances, students are more inclined to help each other, and success and failure on a task are more likely to be attributed to effort rather than to ability. This, in turn, generates conditions that support intrinsic motivation.

Cooperative structures encourage helping and within-group tutoring behaviors, especially when group rewards are based on the performance of all the individual group members. Webb (1985, 1988) has identified the giving or receiving of elaborated explanations as a key factor in student learning within groups, so conditions that favor such activities are desirable. Participation in cooperative learning tends to moderate the positive or negative influence of a student's own high or low performance, tempering both negative and positive self-perceptions resulting from performance, and reducing performance anxiety (Ames, 1984). This can help build both self-esteem and achievement for previously low-achieving students, especially if their group is successful reasonably consistently. Effort attributions are encouraged, partly because the different groups are usually com-

parable in their mix of abilities. Finally, Ames (1984) and Johnson and Johnson (1985) presented evidence that learning in a cooperative group is more enjoyable for most students than learning individually, and that this tends to enhance intrinsic motivation for learning.

Conclusions and Recommendations for Educational Practice

This review began with a caution about the dangers of overgeneralization in educational research. In stating the following conclusions and recommendations, therefore, I must stress that they are not likely to apply in all situations or with all students. Instead, they represent simplifications that appear likely to benefit the greatest proportion of students, and in particular to provide more favorable learning conditions for the weaker students. Many of the specific points draw support from several of the areas of research reviewed earlier, thus increasing the confidence which I have in them.

Importance of evaluation. Classroom evaluation affects students in many different ways. For instance, it guides their judgment of what is important to learn, affects their motivation and self-perceptions of competence, structures their approaches to and timing of personal study (e.g., spaced practice), consolidates learning, and affects the development of enduring learning strategies and skills. It appears to be one of the most potent forces influencing education. Accordingly, it deserves very careful planning and considerable investment of time from educators. Many of the skills and attitudes that are goals of education take years to develop, and their development can be undermined by lack of consistent support for them in the educational experiences of the students (see Howe, 1987; Meyers, 1986).

Classroom evaluation currently appears to receive less thought than most other aspects of education. Its power to affect students is not widely perceived or discussed. A more professional approach to evaluation would demand regular and thoughtful analysis by teachers of their personal evaluation practices, greater use of peer review procedures, and considerable attention to the establishment of more consistent progressions of expectations and criteria within and among educational institutions.

Importance of deep learning. All too often, classroom evaluation places heavy emphasis on the recall or recognition of comparatively isolated pieces of information to which the students have earlier been exposed. This encourages surface (memorizing) approaches to learning. Many of these details have at best only temporary relevance to the students, either because the area studied does not relate to their later activities or interests, or because the details are superceded by new information or developments. Further, it has been repeatedly demonstrated that isolated details are especially readily forgotten, and that information is remembered better and is more useable if students learn it within a broader framework of meaningful interrelationships and understanding. Finally, the knowledge that students accumulate during schooling may be less important than the learning skills and habits they develop, which can help them grow and adapt to new needs and experiences throughout their lifetime. This is increasingly true as modern technology makes factual information available very flexibly and quickly (Rothkopf, 1988, p. 279).

For all these reasons, there is a need to make deep learning a central goal of education, and to foster development of this goal through the evaluation of students

(see also Bloom, 1986; Bok, 1986; Cronbach, 1988; Lowell, 1926; Whitehead, 1929). This requires that we place emphasis on understanding, transfer of learning to untaught problems or situations, and other thinking skills, evaluating the development of these skills through tasks that clearly must involve more than recognition or recall.

These skills take time to develop, however, and are particularly difficult for some students (Lohman, 1986; Thomas, Iventosch, & Rohwer, 1987), so it is important that they be given steadily increasing emphasis from the earliest years of schooling. By the time students are in the upper grade levels or in college, there is a good case for arguing that factual knowledge should be subsumed under higher level objectives, so that students are expected to use factual knowledge in solving a problem or carrying out a process, but are not tested directly on their ability to recall the information.

Evaluation to assist learning. Too much emphasis has been placed on the grading function of evaluation, and too little on its role in assisting students to learn. The integral role of evaluation in teaching and learning needs to be grasped, and its certification function placed in proper perspective. It is hard to see any justification before the final year or so of high school for placing much emphasis on using classroom evaluation for normative grading of student achievement, given the evidence reviewed here that normative grading (with the social comparison and interstudent competition that accompany it) produces undesirable consequences for most students.

These undesirable effects include reduction of intrinsic motivation, debilitating evaluation anxiety, ability attributions for success and failure that undermine student effort, lowered self-efficacy for learning in the weaker students, reduced use and effectiveness of feedback to improve learning, and poorer social relationships among the students. Grading on a fixed curve is especially inappropriate because it emphasizes particularly strongly a comparative approach to grading. Strong emphasis on the grading function of evaluation has also led to overuse of features normally associated with standardized testing, such as very formal testing conditions, speeded tests with strict time limits, a restricted range of item types, and emphasis on the overall score rather than what can be learned about strengths and weaknesses. These may be appropriate in psychological testing, but are rarely appropriate in educational testing (Wood, 1986).

Much of the evaluation activity in education might more profitably be directed solely to giving useful feedback to students, whereas the less frequent evaluations for summative purposes should focus on describing what students can or can't do (i.e., should be criterion referenced). The likely small reduction in reliability associated with counting fewer evaluations in the summative evaluation would be a modest penalty to pay for the benefits described above and the improved validity associated with greater emphasis on final competence (rather than on the mistakes made along the way).

Effective feedback. There are several ways in which the effectiveness of feedback could be enhanced. First, feedback is most effective if it focuses students' attention on their progress in mastering educational tasks. Such emphasis on personal progress enhances self-efficacy, encourages effort attributions, and reduces attention to social comparison. The approach that leads to the most valuable feedback is nicely captured by Easley and Zwoyer (1975):

> If you can both listen to children and accept their answers not as things to just be judged right or wrong but as pieces of information which may reveal what the child is thinking you will have taken a giant step toward becoming a master teacher rather than merely a disseminator of information. (p. 25)

Second, feedback should take place while it is still clearly relevant. This usually implies that it should be provided soon after a task is completed, and that the student should be given opportunities subsequently to demonstrate learning from the feedback. One of the strengths of mastery learning approaches is the emphasis on feedback and subsequent opportunities to correct deficiencies without penalty for the earlier failure.

Third, feedback should be specific and related to need. Simple knowledge of results should be provided consistently (directly or implicitly), with more detailed feedback only where necessary to help the student work through misconceptions or other weaknesses in performance. Praise should be used sparingly and where used should be task specific, whereas criticism (other than simply identifying deficiencies) is usually counterproductive.

Benefits of cooperation. Cooperative learning approaches can be effective in facilitating student learning and motivation and in developing good interpersonal skills and relationships. They are particularly appropriate for more complex tasks where the different perspectives and skills of group members can complement each other.

Approaches that encourage active engagement of all individuals and that stimulate helping behaviors within groups are most desirable. Groups may work together on a group product, but it is also desirable to include some evaluation of the learning of the individual members in the overall evaluation of the achievements of the group.

One of the benefits of cooperative learning is likely to be enhanced development of valuable peer and self-evaluation skills (see Boyd & Cowan, 1985, Johnston & Winograd, 1985), because there is an incentive for groups to monitor their own progress. When normative grading is de-emphasized, cooperative learning is predictably more easy to establish.

Setting standards. Research has repeatedly demonstrated that students achieve most and gain most on key motivational variables when evaluation standards are high but attainable. In many teaching situations this is not possible if all students are working simultaneously on the same tasks and trying to meet the same standards. Under such circumstances, some students will probably not be challenged, whereas others may find the standards unattainable (see Bennett, 1988, p. 26).

To optimize learning outcomes, several alternative approaches are possible. Standards and/or tasks may be set for individual students, or considerable flexibility in learning pathways provided (e.g., mastery learning approaches), or cooperative learning may be used to reduce pressure on individuals and compensate for individual strengths and weaknesses. Weaker students may benefit from identification of more attainable intermediate goals, thus making possible the pattern of repeated successes that leads to improved self-efficacy. Requirements and criteria should be made very clear before an important task is attempted (Anderson & Armbruster, 1984; Natriello, 1987), to avoid misdirected effort and increased evaluation anxiety.

Frequency of evaluation. Students should be given regular opportunities to practice and use the skills and knowledge that are the goals of the program, and to obtain feedback on their performance. Such evaluation fosters active learning, consolidation of learning, and if appropriately arranged can also provide the retention benefits associated with spaced practice. Much of this evaluation can be quite informal, however, and certainly does not need to be conducted under test-like conditions. For higher level outcomes, in particular, it seems likely that too much formal evaluation may be as bad as too little because conceptual understanding and skills do not develop overnight.

Selection of evaluation tasks. The nature and format of evaluation tasks should be selected to suit the goals that are being assessed. In most courses this will lead to substantial variety in tasks, with benefits in versatility of approach and development of transfer skills (Elton, 1982). If it is not inconsistent with program objectives, students could be given some choice of tasks to be attempted. This stimulates and takes advantage of intrinsic motivation, and helps provide suitable challenges for all students.

What is evaluated. The most vital of all the messages emerging from this review is that as educators we must ensure that we give appropriate emphasis in our evaluations to the skills, knowledge, and attitudes that we perceive to be most important. Some of these important outcomes may be hard to evaluate, but it is important that we find ways to assess them. Cross (1987) sums up this point very clearly:

> It serves no useful purpose to lower our educational aspirations because we cannot yet measure what we think is important to teach. Quite the contrary, measurement and assessment will have to rise to the challenge of our educational aspirations. (p. 6)

Concluding remarks. This review has taken a multidimensional look at the impact of classroom evaluation on students. Although the research reviewed is diverse both in focus and in perspective, considerable convergence emerges in the implications of the research findings for effective use of classroom evaluation. My hope is that this review will "help people use their heads" (Cronbach, 1975) in thinking about classroom evaluation, thus enhancing the professionalism and the effectiveness of this important component of teaching and learning.

References

Abbott, R. D., & Falstrom, P. (1977). Frequent testing and personalized systems of instruction. *Contemporary Educational Psychology, 2,* 251–257.

Ames, C. (1984). Competitive, cooperative, and individualistic goal structures: A cognitive-motivational analysis. In R. E. Ames & C. Ames (Eds.), *Research on motivation in education: Vol. 1. Student motivation.* New York: Academic Press.

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 9, pp. 89–132). New York: Academic Press.

Anderson, T. H., & Armbruster, B. B. (1984). Studying. In P. D. Pearson (Ed.), *Handbook of reading research.* New York: Longman.

Astin, A. W. (1987). Competition or cooperation? *Change, 19*(5), 12–19.

Ball, D. W., et al. (1986). Level of teacher objectives and their classroom tests: Match or mismatch. *Journal of Social Studies Research, 10*(2), 27–31.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37,* 122–147.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41,* 586–598.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1987, April). *The impact of peekability on feedback effects.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1988). *Effects of frequent classroom testing.* Unpublished manuscript, University of Michigan.

Becker, H. S., Geer, B., & Hughes, E. C. (1968). *Making the grade: The academic side of college life.* New York: Wiley.

Beeson, R. O. (1973). Immediate knowledge of results and test performance. *Journal of Educational Research, 66,* 224–226.

Bennett, N. (1988). The effective primary-school teacher—The search for a theory of pedagogy. *Teaching and Teacher Education, 4,* 19–30.

Bergland, G. W. (1969). The effect of knowledge of results on retention. *Psychology in the Schools, 6,* 420–421.

Biggs, J. B. (1973). Study behaviour and performance in objective and essay formats. *Australian Journal of Education, 17,* 157–167.

Biggs, J. B. (1978). Individual and group differences in study processes. *British Journal of Educational Psychology, 48,* 266–279.

Bjork, R. A. (1979). Information-processing analysis of college teaching. *Educational Psychologist, 14,* 15–23.

Black, P. J. (1968). University examinations. *Physics Education, 3,* 93–101.

Block, J. H., & Anderson, L. W. (1975). *Mastery learning in classroom instruction.* New York: Macmillan.

Bloom, B. S. (Ed.). (1956). *A taxonomy of educational objectives: Handbook I, the cognitive domain.* New York: Longman.

Bloom, B. S. (1986). Ralph Tyler's impact on evaluation theory and practice. *Journal of Thought, 21,* 36–46.

Bok, D. (1986). Toward higher learning. *Change, 18*(6), 18–27.

Boniface, D. (1985). Candidates' use of notes and textbooks during an open-book examination. *Educational Research, 27,* 201–209.

Boyd, H., & Cowan, J. (1985). A case for self-assessment based on recent studies of student learning. *Assessment and Evaluation in Higher Education, 10,* 225–235.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–374). New York: Macmillan.

Broudy, H. S. (1988). *The uses of schooling.* New York: Routledge.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). In P. H. Mussen (Ed.), *Handbook of child psychology* (Vol. 3, pp. 77–166). New York: Wiley.

Buckwalter, J. A., Schumacher, R., Albright, J. P., & Cooper, R. R. (1981). Use of an educational taxonomy for evaluation of cognitive performance. *Journal of Medical Education, 56,* 115–121.

Carrier, C. A., & Fautsch-Partridge, T. (1981). Levels of questions: A framework for the exploration of processing activities. *Contemporary Educational Psychology, 6,* 365–382.

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist, 19,* 108–120.

Cole, N. S. (1986). Future directions for educational achievement and ability testing. In B. S.

Plake & J. C. Witt (Eds.), *Buros-Nebraska symposium on measurement and testing: Vol. 2. The future of testing.* Hillsdale, NJ: Erlbaum.

Condry, J. C., & Chambers, J. (1978). Intrinsic motivation and the process of learning. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward: New perspectives on the psychology of human motivation.* Hillsdale, NJ: Erlbaum.

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18,* 88–108.

Corno, L., & Rohrkemper, M. M. (1985). The intrinsic motivation to learn in classrooms. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol. 2. The classroom milieu.* New York: Academic Press.

Cotton, J. L., & Cook, M. S. (1982). Meta-analysis and the effects of various reward systems: Some different conclusions from Johnson et al. *Psychological Bulletin, 92,* 176–183.

Covington, M. V. (1984). The motive for self-worth. In R. E. Ames & C. Ames (Eds.), *Research on motivation in education: Vol. 1. Student motivation.* New York: Academic Press.

Covington, M. V. (1985). Strategic thinking and the fear of failure. In J. W. Segal, S. F. Chipman, & R. Glaser (Eds.), *Thinking and learning skills: Vol. 1. Relating instruction to research.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Crooks, T. J. (1982, March). *Generalization in educational research: Through a glass darkly.* Paper presented at the Annual Meeting of the American Educational Research Association, New York. (ERIC Document Reproduction Service No. ED 220 498)

Crooks, T. J., & Collins, E. A. G. (1986). What do first year university examinations assess? *New Zealand Journal of Educational Studies, 21,* 123–132.

Crooks, T. J., & Mahalski, P. A. (1986). Relationships among assessment practices, study methods, and grades obtained. In J. Jones & M. Horsburgh (Eds.), *Research and development in higher education: Vol. 8.* Sydney, Australia: Higher Education Research and Development Society of Australasia.

Cross, K. P. (1987). Teaching for learning. *AAHE Bulletin, 39*(8), 3–7.

Dahlgren, L. O. (1978). Students' conceptions of subject matter: An aspect of learning and teaching in higher education. *Studies in Higher Education, 3,* 25–35.

deCharms, R. (1976). *Enhancing motivation: Change in the classroom.* New York: Irvington.

deCharms, R. (1983). Intrinsic motivation, peer tutoring, and cooperative learning. In J. M. Levine & M. C. Wang (Eds.), *Teacher and student perceptions: Implications for learning* (pp. 391–398). Hillsdale, NJ: Erlbaum.

Deci, E. L. (1975). *Intrinsic motivation and self-determination in human behavior.* New York: Irvington.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self determination in human behavior.* New York: Plenum.

Dempster, F. N. (1987). Time and the production of classroom learning: Discerning implications from basic research. *Educational Psychologist, 22,* 1–21.

Deutsch, M. (1979). Education and distributive justice: Some reflections on grading systems. *American Psychologist, 34,* 391–401.

Dillon, J. T. (1982). Cognitive correspondence between question/statement and response. *American Educational Research Journal, 19,* 540–551.

DiSibio, M. (1982). Memory for connected discourse: A constructivist view. *Review of Eduational Research, 52,* 149–174.

Dorr-Bremme, D. W., & Herman, J. (1986). *Assessing school achievement: A profile of classroom practices.* Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53,* 159–199.

Doyle, W. (1986) Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., 392–431). New York: Macmillan.

Driscoll, M. P. (1986). The relationship between grading standards and achievement: A new perspective. *Journal of Research and Development in Education, 19*(3), 13–17.

Duchastel, P. C. (1979a). Adjunct question effects and experimental constraints. Occasional Paper 1, American College, Bryn Mawr, PA. (ERIC Document Reproduction Service No. ED 216 312)

Duchastel, P. C. (1979b). Retention of prose materials: The effect of testing. *Journal of Educational Research, 72,* 299–300.

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology, 6,* 217–226.

Duchastel, P. C. (1983). Interpreting adjunct question research: Processes and ecological validity. *Human Learning, 2,* 1–5.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75,* 309–313.

Dweck, C. S., & Bempechat, J. (1983). Children's theories of intelligence: Consequences for learning. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom.* Hillsdale, NJ: Erlbaum.

Dweck, C. S., & Elliott, E. S. (1983). Achievement motivation. In P. H. Mussen (Ed.), *Handbook of child psychology* (Vol. 4, pp. 643–691). New York: Holt, Rinehart and Winston.

d'Ydewalle, G., Swerts, A., & De Corte, E. (1983). Study time and test performance as a function of test expectations. *Contemporary Educational Psychology, 8,* 55–67.

Easley, J. A., & Zwoyer, R. E. (1975). Teaching by listening—Toward a new day in math classes. *Contemporary Education, 47,* 19–25.

Ebel, R. L. (1982). Proposed solution to two problems of test construction. *Journal of Educational Measurement, 19,* 267–278.

Eccles, J. (1983). Expectancies, values and academic behavior. In J. T. Spence (Ed.), *Academic and achievement motives.* San Francisco: Freeman.

Elton, L. R. B. (1982). Assessment for learning. In D. Bligh (Ed.), *Professionalism and flexibility for learning.* Guildford, Surrey, England: Society for Research into Higher Education.

Elton, L. R. B., & Laurillard, D. M. (1979). Trends in research on student learning. *Studies in Higher Education, 4,* 87–102.

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

Entwistle, N. J., & Ramsden, P. (1983). *Understanding student learning.* London: Croom Helm.

Ericksen, S. C. (1983). Private measures of good teaching. *Teaching of Psychology, 10,* 133–136.

Fennessy, D. (1982, July). *Primary teachers' assessment practices: Some implications for teacher training.* Paper presented at the 12th annual conference of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia. (ERIC Document Reproduction Service No. ED 229 346)

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *New directions for testing and measurement: Vol. 19. Testing in the schools.* San Francisco: Jossey-Bass.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80,* 179–183.

Ford, N. (1981). Recent approaches to the study and teaching of "effective learning" in higher education. *Review of Educational Research, 51,* 345–377.

Francis, J. (1982). A case for open-book examinations. *Educational Review, 34,* 13–26.
Fredericksen, N. (1984a). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193–202.
Fredericksen, N. (1984b). Implications of cognitive theory for instruction in problem solving. *Review of Educational Research, 54,* 363–407.
Friedman, H. (1987). Repeat examinations in introductory statistics. *Teaching of Psychology, 14,* 20–23.
Gagne, R. M. (1977). *The conditions of learning* (3rd ed.). New York: Holt, Rinehart and Winston.
Gagne, R. M., Briggs, L. J., & Wager, W. W. (1988). *Principles of instructional design.* New York: Holt, Rinehart and Winston.
Gall, M. (1984). Synthesis of research on teachers' questioning. *Educational Leadership, 42*(3), 40–47.
Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement, 17,* 45–50.
Glaser, R. (1985, November). *The integration of instruction and testing.* Paper presented at the E.T.S. Invitational Conference on the Redesign of Testing for the 21st Century, New York.
Goslin, D. A. (1967). *Teachers and testing* (2nd ed.). New York: Russell Sage Foundation.
Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77,* 244–248.
Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79,* 96–100.
Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice, 4*(1), 15–18.
Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership, 43*(8), 73–80.
Guza, D. S., & McLaughlin, T. F. (1987). A comparison of daily and weekly testing on student spelling performance. *Journal of Educational Research, 80,* 373–376.
Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55,* 23–46.
Haertel, E. (1986, April). *Choosing and using classroom tests: Teachers' perspectives on assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
Hakstian, R. (1971). The effects of type of examination anticipated on test preparation and performance. *Journal of Educational Research, 64,* 319–324.
Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56,* 212–242.
Hamilton, R. J. (1985). A framework for the evaluation of the effectiveness of adjunct questions and objectives. *Review of Educational Research, 55,* 47–85.
Harter, S. (1985). Competence as a dimension of self-evaluation: Toward a comprehensive model of self-worth. In R. Leahy (Ed.), *The development of the self.* New York: Academic Press.
Hill, K. T. (1984). Debilitating motivation and testing: A major educational problem—Possible solutions and policy applications. In R. E. Ames & C. Ames (Eds.), *Research on motivation in education: Vol. 1. Student motivation.* New York: Academic Press.
Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *Elementary School Journal, 85,* 105–126.
Howe, M. J. A. (1987). Using cognitive psychology to help students learn how to learn. In J. T. E. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student learning: Research in education and cognitive psychology.* Milton Keynes, England: Open University Press & Society for Research into Higher Education.

Hughes, B., Sullivan, H. J., & Mosley, M. L. (1985). External evaluation, task difficulty, and continuing motivation. *Journal of Educational Research, 78,* 210–215.

Hunkins, F. P. (1969). Effects of analysis and evaluation questions on various levels of achievement. *Journal of Experimental Education, 38*(2), 45–58.

Ingenkamp, K. (1986). The possible effects of various reporting methods on learning outcomes. *Studies in Educational Evaluation, 12,* 341–350.

Johnson, D. W., & Johnson, R. T. (1985). Motivational processes in cooperative, competitive, and individualistic learning situations. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol. 2. The classroom milieu.* New York: Academic Press.

Johnson, D. W., Johnson, R. T., & Maruyama, G. (1983). Interdependence and interpersonal attraction among heterogeneous and homogeneous individuals: A theoretical formulation and a meta-analysis of the research. *Review of Educational Research, 53,* 5–54.

Johnson, D. W., Maruyama, G., and Johnson, R. T. (1982). Separating ideology from currently available data: A reply to Cotton and Cook and McGlynn. *Psychological Bulletin, 92,* 186–192.

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47–62.

Johnston, P. H., & Winograd, P. N. (1985). Passive failure in reading. *Journal of Reading Behavior, 17,* 279–301.

Jones, H. E. (1923). Experimental studies of college teaching: The effect of examination on permanence of learning. *Archives of Psychology, 10,* 1–70.

Karraker, R. J. (1967). Knowledge of results and incorrect recall of plausible multiple-choice alternatives. *Journal of Educational Psychology, 58,* 11–14.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). *The effects of standardized testing.* Boston: Kluwer-Nijhoff.

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 383–434). Hillsdale, NJ: Erlbaum.

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology, 25,* 427–436.

Kifer, E. (1977). The impact of success and failure on the learner. *Evaluation in Education: International Progress, 1,* 281–359.

Kirkland, M. C. (1971). The effects of tests on students and schools. *Review of Educational Research, 41,* 303–350.

Klinzing, G., Klinzing-Eurich, G., & Tisher, R. P. (1985). Higher cognitive behaviours in classroom discourse: Congruencies between teachers' questions and pupils' responses. *Australian Journal of Education, 29,* 63–75.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47,* 211–232.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple choice tests. *Journal of Educational Psychology, 63,* 505–512.

Kulhavy, R. W., Dyer, J. W., & Silver, L. (1975). The effects of note-taking and test expectancy on the learning of text material. *Journal of Educational Research, 68,* 363–365.

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

Kulik, J. A., & Kulik, C-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79–97.

Kumar, V. K., Rabinsky, L., & Pandey, T. N. (1979). Test mode, test instructions, and retention. *Contemporary Educational Psychology, 4,* 211–218.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology, 67,* 259–266.

Laurillard, D. (1979). The processes of student learning. *Higher Education, 8,* 395–409.

Laurillard, D. (1984). Learning from problem-solving. In F. Marton, D. J. Hounsell, & N. J.

Entwistle (Eds.), *The experience of learning.* Edinburgh: Scottish Academic Press.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology, 28,* 129–137.

Levin, J. R. (1982). Pictures as prose-learning devices. In A. Flamme & W. Kintsch (Eds.), *Advances in psychology: Vol. 8. Discourse processing.* Amsterdam: North-Holland.

Lew, M., Mesch, D., Johnson, D. W., & Johnson, R. (1986). Positive interdependence, academic and collaborative-skills group contingencies, and isolated students. *American Educational Research Journal, 23,* 476–488.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20,* 179–189.

Lohman, D. F. (1986). Predicting mathemathanic effects in the teaching of higher-order thinking skills. *Educational Psychologist, 21,* 191–208.

Lowell, A. L. (1926). The art of examination. *Atlantic Monthly, 137,* 58–66.

Madaus, G. F., & Airasian, P. W. (1977). Issues in evaluating student outcomes in competency-based graduation programs. *Journal of Research and Development in Education, 10,* 79–91.

Madaus, G. F., & McDonagh, J. T. (1979). Minimum competency testing: Unexamined assumptions and unexplored negative outcomes. In R. T. Lennon (Ed.), *Impactive changes in measurement,* New Directions for Testing and Measurement, No. 3. San Francisco: Jossey-Bass.

Maehr, M. L. (1976). Continuing motivation: An analysis of a seldom considered educational outcome. *Review of Educational Research, 46,* 443–462.

Maehr, M. L. (1983). Doing well in science: Why Johnny no longer excels; why Sarah never did. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom.* Hillsdale, NJ: Erlbaum.

Maehr, M. L., & Stallings, W. M. (1972). Freedom from external evaluation. *Child Development, 43,* 177–185.

Martin, E., & Ramsden, P. (1987). Learning skills and skill in learning. In J. T. E. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student learning: Research in education and cognitive psychology.* Milton Keynes, England: Open University Press & Society for Research into Higher Education.

Marton, F., Hounsell, D. J., & Entwistle, N. J. (Eds.). (1984). *The experience of learning.* Edinburgh: Scottish Academic Press.

Marton, F., & Säljö, R. (1976a). On qualitative differences in learning: 1. Outcome and process. *British Journal of Educational Psychology, 46,* 4–11.

Marton, F., & Säljö, R. (1976b). On qualitative differences in learning: 2. Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology, 46,* 115–127.

Mathews, J. (1980). *The uses of objective tests.* Teaching in Higher Education Series, No. 9. England: Lancaster University. (ERIC Document Reproduction Service No. ED 230 106)

Mayer, R. E. (1975). Forward transfer of different reading strategies evoked by testlike events in mathematics text. *Journal of Educational Psychology, 67,* 165–169.

McCombs, B. L. (1984). Processes and skills underlying continuing intrinsic motivation to learn: Toward a definition of motivational skills training interventions. *Educational Psychologist, 19,* 199–218.

McGlynn, R. P. (1982). A comment on the meta-analysis of goal structures. *Psychological Bulletin, 92,* 184–185.

McKeachie, W. J. (1974). The decline and fall of the laws of learning. *Educational Researcher, 3,* 7–11.

McKeachie, W. J. (1984). Does anxiety disrupt information processing or does poor information processing lead to anxiety. *International Review of Applied Psychology, 33,* 187–203.

McKeachie, W. J., Pintrich, P. R., Lin, Y., & Smith, D. A. F. (1986). *Teaching and learning in the college classroom: A review of the research literature.* Ann Arbor, Michigan: National Center for Research to Improve Postsecondary Teaching and Learning.

McPartland, J. M. (1987, April). *Changing testing and grading practices to improve student motivation and teacher-student relationships: Designs for research to evaluate new ideas for departmental exams and progress grades.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Medley, D. M. (1979). The effectiveness of teachers. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching.* Berkeley, CA: McCutchan.

Messick, S. (1984a). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215–237.

Messick, S. (1984b). Abilities and knowledge in educational achievement testing: The assessment of dynamic cognitive structures. In B. S. Plake (Ed.), *Buros-Nebraska symposium on measurement and testing: Vol. 1. Social and technical issues in testing: Implications for test construction and usage.* Hillsdale, NJ: Erlbaum.

Meyer, G. (1934). An experimental study of the old and new types of examination: 1. The effect of the examination set on memory. *Journal of Educational Psychology, 25,* 641–661.

Meyer, G. (1935). An experimental study of the old and new types of examination: 2. Methods of study. *Journal of Educational Psychology, 26,* 30–40.

Meyers, C. (1986). *Teaching students to think critically.* San Francisco: Jossey-Bass.

Miller, C. M. L., & Parlett, M. (1974). *Up to the mark: A study of the examination game.* London: Society for Research into Higher Education.

Miller, G. E. (1976). Continuous assessment. *Medical Education, 10,* 81–86.

Miller, G. E. (1978). 'Teaching and learning in medical school' revisited. *Medical Education, 12,* Supplement, 120–125.

Mills, S. R., Rice, C. T., Berliner, D. C., & Rousseau, E. W. (1980). The correspondence between teacher questions and student answers in classroom discourse. *Journal of Experimental Education, 48,* 194–204.

Milton, O. (1982). *Will that be on the final?* Springfield, IL: Charles C. Thomas.

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

Natriello, G., & Dornbusch, S. M. (1984). *Teacher evaluative standards and student effort.* New York: Longman.

Naveh-Benjamin, M., McKeachie, W. J., & Lin, Y-G. (1987). Two types of test-anxious students: Support for an information processing model. *Journal of Educational Psychology, 79,* 131–136.

Newble, D. I., & Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education, 17,* 25–31.

Nicholls, J. G. (1983). Conceptions of ability and achievement motivation: A theory and its implications for education. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom.* Hillsdale, NJ: Erlbaum.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review, 91,* 328–346.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74,* 18–22.

O'Neill, M., Razor, R. A., & Bartz, W. R. (1976). Immediate retention of objective test answers as a function of feedback complexity. *Journal of Educational Research, 70,* 72–75.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology, 49,* 173–181.

Paris, S. G. (1988, April). *Fusing skill and will in children's learning and schooling.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Paris, S. G., & Cross, D. R. (1983). Ordinary learning: Pragmatic connections among children's

beliefs, motives, and actions. In J. Bisanz, G. Bisanz, & R. Kail (Eds.), *Learning in children* (pp. 137–169). New York: Springer-Verlag.

Perry, W. F. (1970). *Forms of intellectual and ethical development in the college years: A scheme.* New York: Holt, Rinehart and Winston.

Phye, G. D. (1979). The processing of informative feedback about multiple-choice test performance. *Contemporary Educational Psychology, 4,* 381–394.

Pressley, M., Levin, J. R., & Chatala, E. S. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior, 23,* 270–288.

Quellmalz, E. S. (1985). Needed: Better methods for testing higher-order thinking skills. *Educational Leadership, 43*(2), 29–35.

Ramsden, P. (1984). The context of learning. In F. Marton, D. J. Hounsell, & N. J. Entwistle (Eds.), *The experience of learning.* Edinburgh: Scottish Academic Press.

Ramsden, P. (1985). Student learning research: Retrospect and prospect. *Higher Education Research and Development, 4,* 51–69.

Ramsden, P., Beswick, D., & Bowden, J. (1987). Learning processes and learning skills. In J. T. E. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student learning: Research in education and cognitive psychology.* Milton Keynes, England: Open University Press & Society for Research into Higher Education.

Ramsden, P., & Entwistle, N. J. (1981). Effects of academic departments on students' approaches to studying. *British Journal of Educational Psychology, 51,* 368–383.

Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51,* 237–245.

Rickards, J. P., & Friedman, F. (1978). The encoding versus the external storage hypothesis in notetaking. *Contemporary Educational Psychology, 3,* 136–143.

Rinchuse, D. J., & Zullo, J. (1986). The cognitive level demands of a dental school's predoctoral didactic examinations. *Journal of Dental Education, 50,* 167–171.

Rogers, E. M. (1969). Examinations: Powerful agents for good or ill in teaching. *American Journal of Physics, 37,* 954–962.

Rohm, R. A., Sparzo, F. J., & Bennett, C. M. (1986). College student performance under repeated testing and cumulative testing conditions: Reports on five studies. *Journal of Educational Research, 80,* 99–104.

Rohwer, W. D., & Thomas, J. W. (1987). The role of mnemonic strategies in study effectiveness: Theories, individual differences, and applications. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes.* New York: Springer-Verlag.

Rosenholtz, S. J., & Simpson, C. (1984). Classroom organization and student stratification. *Elementary School Journal, 85,* 21–37.

Rosenshine, B. (1979). Content, time, and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching.* Berkeley, CA: McCutchan.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 376–391). New York: Macmillan.

Rosswork, S. G. (1977). Goal setting: The effects of an academic task with varying magnitudes of incentive. *Journal of Educational Psychology, 69,* 710–715.

Rothkopf, E. Z. (1988). Perspectives on study skills training in a realistic instructional economy. In C. E. Weinstein, E. T. Goetz, & P. A. Alexander (Eds.), *Learning and study strategies: Issues in assessment, instruction, and evaluation.* San Diego, CA: Academic Press.

Rowe, D. W. (1986). Does research support the use of "purpose questions" on reading comprehension tests? *Journal of Educational Measurement, 23,* 43–55.

Rowe, M. B. (1986). Wait time: Slowing down may be a way of speeding up! *Journal of Teacher Education, 37*(1), 43–50.

Rudman, H. E., Kelley, J. L., Wanous, D. S., Mehrens, W. A., Clark, C. M., & Porter, A. C. (1980). *Integrating assessment with instruction: A review (1922–1980)* (Research Series No.

75). East Lansing, MI: Michigan State University, Institute for Research on Teaching. (ERIC Document Reproduction Service No. ED 189 136).

Ryan, R. M., Connell, J. P., & Deci, E. L. (1985). A motivational analysis of self-determination and self-regulation in education. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol. 2. The classroom milieu.* New York: Academic Press.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education, 54,* 60–79.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan, 62,* 631–634.

Samson, G. E., Strykowski, B., Weinstein, T., & Walberg, H. J. (1987). The effects of teacher questioning levels on student achievement: A quantitative synthesis. *Journal of Educational Research, 80,* 290–295.

Sarason, I. R. (1980). *Test anxiety: Theory and applications.* Hillsdale, NJ: Erlbaum.

Sassenrath, J. M., & Garverick, C. M. (1965). Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology, 56,* 259–263.

Sax, G., & Collet, L. S. (1968). An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement, 5,* 169–173.

Schmeck, R. R. (1983). Learning styles of college students. In R. Dillon & R. R. Schmeck, *Individual differences in cognition.* New York: Academic Press.

Schmeck, R. R. (1988). Individual differences and learning strategies. In C. E. Weinstein, E. T. Goetz, & P. A. Alexander (Eds.), *Learning and study strategies: Issues in assessment, instruction, and evaluation.* San Diego, CA: Academic Press.

Schunk, D. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist, 19,* 48–58.

Schunk, D. (1985). Self-efficacy and classroom learning. *Psychology in the Schools, 22,* 208–223.

Shavelson, R. J., Berliner, D. C., Ravitch, M. M., & Loeding, D. (1974). Effects of position and type of question on learning from prose material: Interaction of treatments with individual differences. *Journal of Educational Psychology, 66,* 40–48.

Shulman, L. S. (1980). Test design: A view from practice. In E. L. Baker & E. S. Quellmaltz (Eds.), *Educational testing and evaluation.* Beverly Hills, CA: Sage.

Slavin, R. E. (1978). Separating incentives, feedback, and evaluation: Toward a more effective classroom system. *Educational Psychologist, 13,* 97–100.

Slavin, R. E. (1983a). *Cooperative learning.* New York: Longman.

Slavin, R. E. (1983b). When does cooperative learning increase student achievement? *Psychological Bulletin, 94,* 429–445.

Slavin, R. E. (1984). Students motivating students to excel: Cooperative incentives, cooperative tasks, and student achievement. *Elementary School Journal, 85,* 53–63.

Slavin, R. E. (1987). Developmental and motivational perspectives on cooperative learning: A reconciliation. *Child Development, 58,* 1161–1167.

Snyder, B. R. (1971). *The hidden curriculum.* Cambridge, MA: M.I.T. Press.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656.

Stice, J. E. (1987). Learning how to think: Being earnest is important but it's not enough. In J. E. Stice (Ed.), *New directions for teaching and learning: Vol. 30. Developing critical thinking and problem-solving abilities.* San Francisco: Jossey-Bass.

Stiggins, R. J. (1985). Improving assessment where it means the most: In the classroom. *Educational Leadership, 43*(2), 69–74.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22,* 271–286.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5*(2), 5–17.

Stiggins, R. J., Griswold, M., Green, K. R., & associates (1988, April). *Measuring thinking*

skills through classroom assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Strang, H. R., & Rust, J. O. (1973). The effect of immediate knowledge of results and task definition on multiple-choice answering. *Journal of Experimental Education, 42,* 77–80.

Svensson, L. (1977). On qualitative differences in learning: 3. Study skill and learning. *British Journal of Educational Psychology, 47,* 233–243.

Terry, P. W. (1933). How students review for objective and essay tests. *Elementary School Journal, 33,* 592–603.

Thomas, J. W., Iventosch, L., & Rohwer, W. D. (1987). Relationships among student characteristics, study activities, and achievement as a function of course characteristics. *Contemporary Educational Psychology, 12,* 344–364.

Thomas, J. W., & Rohwer, W. D. (1986). Academic studying: the role of learning strategies. *Educational Psychologist, 21,* 19–41.

Thorndike, E. L., & Woodyard, E. (1934). The influence of the relative frequency of success and frustrations upon intellectual achievement. *Journal of Educational Psychology, 25,* 241–250.

Thorndike, R. L. (1969). Helping teachers use tests. *NCME Measurement in Education, 1*(1), 1–4.

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist, 20,* 135–142.

van Rossum, E. J., Diejkers, R., & Hamer, R. (1985). Students' learning conceptions and their interpretation of significant educational concepts. *Higher Education, 14,* 617–641.

van Rossum, E. J., & Schenk, S. M. (1984). The relationship between learning conception, study strategy, and learning outcome. *British Journal of Educational Psychology, 54,* 73–83.

Watkins, D. (1984). Students' perceptions of factors influencing tertiary learning. *Higher Education Research and Development, 3,* 33–50.

Watts, G., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62,* 387–394.

Webb, N. M. (1985). Student interaction and learning in small groups: A research summary. In R. E. Slavin, S. Sharan, S. Kagan, R. Hertz-Lazarowitz, C. Webb, & R. Schmuck (Eds.), *Learning to cooperate, cooperating to learn* (pp. 147–172). New York: Plenum.

Webb, N. M. (1988, April). *Small group problem-solving: Peer interaction and learning.* Invited address at the annual meeting of the American Educational Research Association, New Orleans, LA.

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology, 71,* 3–25.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92,* 548–573.

Weiner, B. (1986). *An attributional theory of motivation and emotion.* New York: Springer-Verlag.

Wexley, K. N., & Thornton, C. L. (1972). Effect of verbal feedback of test results on learning. *Journal of Educational Research, 66,* 119–121.

Whitehead, A. N. (1929). *The aims of education.* New York: Macmillan.

Wilson, J. D. (1981). *Student learning in higher education.* London: Croom Helm.

Winne, P. H. (1979). Experiments relating teachers' use of higher cognitive level questions to student achievement. *Review of Educational Research, 49,* 13–50.

Wittrock, M. C. (1974). Learning as a generative process. Educational Psychologist, 11, 87–95.

Wittrock, M. C. (1979). The cognitive movement in instruction. *Educational Psychologist, 13,* 15–29.

Wittrock, M. C. (1986). Students' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 297–314). New York: Macmillan.

Wood, R. (1986). The agenda for educational measurement. In D. L. Nuttall (Ed.), *Assessing educational achievement.* London: Falmer Press.

Author

TERENCE J. CROOKS, Senior Lecturer, Director, Higher Education Development Centre, University of Otago, P.O. Box 56, Dunedin, New Zealand. *Specializations*: improvement of tertiary education, research design, measurement and evaluation.