EDUCATORS’ HANDBOOK
A RESEARCH PERSPECTIVE

Virginia Richardson-Koehler
Senior Editor

Section Editors
David C. Berliner
Ursula Casanova
Christopher M. Clark
Richard H. Hersh
Lee S. Shulman

Longman
New York & London
All teachers experience a moment of truth, a moment of great joy or bitter disappointment. It is the moment when teachers check their students' understanding. Did they or didn't they understand what was taught to them? If the students appeared to understand what was taught, teachers feel happy and fulfilled. On the other hand, if the students failed to understand what was taught, teachers usually attribute that failure to the students’ lack of preparation, their low ability, poor home conditions, or the difficulty of the test. In other words, teachers take credit for student success but blame student failure on causes other than themselves (Bar-Tal, 1979). This is a very human quality, but it probably is not appropriate professional behavior.

Appropriate professional behavior by teachers requires first the delivery of extremely comprehensible explanations to ensure, as far as is possible, that what is communicated by teachers can be understood by students. After such presentations, teachers need to inquire whether in fact the understanding they wanted to promote actually occurred. These two issues—explaining for understanding and checking for understanding—are the focus of this chapter. First we will investigate ways to provide “good” explanations so that teachers can justifiably take some credit when students demonstrate that they have understood instruction. Then we will discuss both formal and informal ways in which teachers can check for understanding. Teachers cannot judge their own effectiveness, nor can students evaluate their own progress, unless such checking takes place.

EXPLAINING FOR UNDERSTANDING

Research has examined only a few of the many ways in which a teacher can help ensure that a student understands what the teacher is trying to teach. The research suggests that teachers help students learn more and better if they provide some structure for new information, prompt awareness of the students’ own relevant
information, and follow a model for organizing explanations. In addition, some helpful tips from the research can make what teachers teach more easily understood by their students.

Providing Structures for New Knowledge

Telling students in advance about what they are going to learn, what the key points to be mastered are, and what they should know at the end of an instructional episode has been positively related to student achievement. This seems to hold for both short instructional episodes and for those lasting over days or weeks. Classroom observations have shown that teachers do not always remember to do this. They often simply plunge into an instructional unit, forgetting to make explicit to students the aims of that instruction.

But over and above such helpful, structuring, prefatory remarks, teachers can help students understand instruction by providing advance organizers. An advance organizer provides the learner with some rules for organizing a body of apparently unorganized material, or an organizer can take the form of higher-level propositions about the material to be learned (Ausubel, 1968). Organizers provide concepts under which to subsume other concepts. They are brief statements at a relatively high level of abstraction, generality, and inclusiveness compared to the material to be learned. These high-level rules, propositions, and concepts help students to understand new instructional content by providing them with concepts on which to hang or anchor the new ideas. That is, the advance organizer provides some way to structure the new knowledge that the teacher is trying to impart. These hooks or anchoring concepts have been found to help students identify, store, and retrieve information better. They are intended to provide a kind of "ideational scaffolding" for the content of instruction. An organizer is not just an introduction; it provides verbal structures to aid students in fitting new information into their existing knowledge systems. To "understand" is to have integrated new information into one's personal store of knowledge. Understanding is distinguished from memory by this kind of integration.

Prompting Awareness of Relevant Knowledge

Understanding requires the integration of new information with one's existing knowledge. Teachers can help students to understand by making explicit the relationship of the new instructional content to other knowledge possessed by the students. Whether the content is the addition of negative numbers or the study of the Revolutionary War, new content will be understood better by students if it is explicitly related to what they already know. Teachers either have to create the structure for integrating new material with already known material, or they need to prompt the students deliberately to bring to consciousness the relevant knowledge they have about the topic being taught. In the language of modern cognitive psychology this is called "engaging the relevant schema." Schemata are abstract representations of knowledge that we possess about objects and concepts. The schemata a learner brings to a situation may determine what and how much is understood from that situation.

Imagine a section from a geography text about an unfamiliar nation. An adult would bring to bear an elaborate nation schema, which would point to subchema representing generic knowledge about political systems, economics, geography, and climate. Each subchema would have its own infrastructure and interconnect with other subchema at various points. . . . The young reader, on the other hand, may not possess a nation schema adequate to assimilate the text. In the worst case, the material will be gibberish. . . more likely, the young reader will have partly formed schemata that will allow him or her to make sense of the passage, but will not permit the construction of mental representatives of great depth or breadth. (Anderson, Spiro, & Anderson, 1978, p. 439)

Understanding, as the quote makes clear, depends on engaging appropriate schemata. It is now hypothesized that the reason metaphors are useful in instruction ("electric current is like flowing water"; "programming a computer is like instructing a stupid child") is because new information can be incorporated into already existing schema, with only minor changes. Thus whatever ways a teacher can use to tie new ideas to old ones are likely to help students understand instruction better.

A Model for Explaining

When teachers have to explain why it is warmer in summer than in winter, why time differs in New York and San Francisco, and why multiplying two negative numbers produces a positive number, they must construct a very complex communication. Giving an explanation that results in understanding by most students is not a simple task. In part this is because the content of each explanation is so different and because student questions that often trigger such explanations are so unpredictable. Nevertheless, a simple and learnable model for explaining diverse phenomena in ways that foster understanding has been developed:

Step 1. Make sure you understand the question that a student has asked or that you raise in your presentation. Ask yourself what it is that most students who receive the explanation should be concerned about.

Step 2. Identify the "things" (elements, variables, concepts, events) involved in the relationship needing to be explained.

Step 3. State the relationships between the different "things" you identified in step 2.

Step 4. Show how the relationship identified in step 3 is an instance of a more general relationship or principle.

This model fits the description of what it means to come to "understand" something, as described by J. M. Thyne (1963, pp. 127–129) and illustrated here:
more from teachers who said, “Pick up your blue pen. First draw a big blue circle and then draw a little blue circle next to it” than from teachers who said, “In blue, after you have drawn a big circle, draw a little one by it.”

The importance of explicitness in the teacher’s instructions to students about what to do is rather obvious. Less obvious is the importance of explicitness in the teacher’s instructions to students about how to do something. For example, a teacher might introduce an assignment on finding the meaning of words that have prefixes either very simply or by explicitly demonstrating (1) how to divide words into root and prefix, (2) how to determine the meaning of the root, (3) how to determine the meaning of the prefix, and (4) how to put prefix and root together in a meaningful way. Low achievers who did not know how to do this kind of seatwork exercise spent their time copying answers and guessing. They may have understood what was expected of them, but they did not understand how they were to do it. This is really not surprising, given the fact that when teachers’ explanations of seatwork assignments were analyzed, only 1.5% included explicit descriptions of the cognitive processes that the teacher had hoped the students would use. (Communication Quarterly, 1984; the specific work referred to in this section is by L. Anderson, G. Duffy, and L. Roehler of the Institute for Research on Teaching, Michigan State University, and by P. Marland, James Cook University, Townsville, Australia.)

Able students apparently seek out and find the cues emitted by their unexplicit teachers. They note the words that seem to be important to their teachers, the concepts used in summaries and introductions, the ideas put on the chalkboard or in a handout, and similar cues. From these they determine what is worth trying to learn. Less able students are “cue-blind.” They never learn to identify from the myriad bits of information presented what it is that should be understood.

Explicitness in teaching reading comprehension has also been studied, to try to put an end to the often heard student statement, “I read it, but it sure doesn’t make any sense!” Teachers can do a better job of teaching how to comprehend written material by doing the following three things:

1. Focusing on the specific mental processes needed to do the task at hand. (The how to do, the process, is at least as important to communicate to students as the what to do, the product.)

2. Making visible the mental process involved. (For example, the steps toward the solution of a problem can be verbalized, not left silent, allowing a teacher to model the way the relevant mental processes should be employed.)

3. Making instruction cohesive and continuous across lessons. (It is hard to understand what is expected when instruction is disjointed and disconnected.)

When these and some other techniques that were believed to promote understanding were turned into training packages, teachers became more explicit in their instruction. The students of these trained teachers, in comparison to the students of untrained teachers, were more involved in their reading lessons, more conscious of their reading skills and strategies, and able to make more sense out of what they read. The researchers noted, however, how hard it was to change people who are not explicit into people who are explicit. Nevertheless, ways to improve explicitness must be worked at. Teachers who give vague directions to students about what to do and who do not make clear to students how they are to think about a problem are not good instructors, and their faults must be corrected.

The Rule-Example-Rule Technique. When compared to less effective communicators, the explanations of very effective communicators were found to contain many more instances of a pattern of phrases called rule-example-rule. Here is an excerpt from a teacher who was found to be high in effectiveness. This teacher was explaining aspects of Yugoslavian foreign policy.

As a communist but not a Soviet-dominated nation, they want to increase peaceful and friendly relations with other nations in Europe. They want to do this by having more trade, through exchanging ideas, and through more personal contact with other nations in Europe. They believe that better relations must be established between countries in Europe.

In this example, the first and last sentences are rules, and the intervening sentence is an example. The sequence of stating a rule, giving examples, and stating the rule again helps students understand explanations.

Using Examples. Even without being part of a rule-example-rule sequence, the liberal use of examples appears to be beneficial. To confirm the positive effect that using examples has on students understanding is easy. You might try to do your best to provide explanations to a group of students. Then try to redo the presentation to an equivalent group of students, this time inserting two examples to illustrate each concept. Then evaluate your instruction. When this was done as part of a research project, it was found that the students understood more from the revised explanation and rated it more favorably. You will probably get the same result.

Explanatory Links. Rosenshine (1971) has reported that good explainers link their phrases in a special way. They use prepositions and conjunctions that indicate the cause, result, means, or purposes of an event or idea. Examples include because, in order of, if... then, therefore, consequently, and certain uses of since, by, and through. Explanatory links tie ideas together either within or between sentences. Here are four sentences; the first one does not contain an explanatory link, whereas the last three do.

1. The Chinese dominate Bangkok’s economy, and they are a threat.
2. The Chinese dominate Bangkok’s economy; therefore, they are a threat.
3. The Chinese are a threat because they dominate Bangkok’s economy.
4. By dominating Bangkok’s economy, the Chinese became a threat.

Explanatory links probably work because (an explanatory link!) they cue the learner that a relationship is being described, and relationships help tie the ideas together, making them more meaningful and easier to learn.

Verbal Markers of Importance. Verbal markers of importance are cues to students that indicate what they should attend to most in an explanation. The cues seem to tell the learners, “This is the key to it!” Examples of such cues are phrases like Now note this; It is especially important to realize that...; It will help you to understand this better if you remember that...; Now let’s discuss the most crucial point of all, namely, that... From more than a half dozen research studies we have evidence that using such markers in explanations increases students understanding.
CHECKING FOR UNDERSTANDING

Ultimately, after instruction of a few minutes, hours, or weeks, the time comes to check for understanding. For relatively short instructional episodes, and to help teachers determine the pace and form of instruction in longer episodes, teachers use informal methods of checking. These include interpreting the cues emitted by students in the classroom, framing classroom questions and analyzing their answers, and some other relatively informal, usually unplanned, methods to decide if understanding has taken place. Teachers also use some semiformal methods for assessing student understanding. These include the assignment and monitoring of classroom practice and the assignment and analysis of homework. Formal methods to assess student understanding and to assign marks or grades based on that assessment involve the use of tests.

Informal Methods

The most frequently used methods for checking on understanding are the teachers' interpretation of classroom cues, the use of classroom questions, and conferencing.

Classroom Cues. Imagine the following situation. A student receives a good explanation about some phenomenon. The student then opens a booklet and answers questions about the phenomenon, one question and answer per page. A hidden camera is focused on the student. This is repeated for many students. The questions are scored; some are found to be answered correctly, and some are clearly incorrect. What this provides, then, are film clips of students who either did or did not understand something. Now imagine that these film clips are shown to teachers, who are asked to make a simple choice: Did the student get a particular item right or wrong? If teachers have learned to read the cues in a classroom, they should do far better than chance when making these kinds of judgments. Teachers who observe students as they respond to questions that check their understanding ought to be able to interpret correctly the dozens of nonverbal cues emitted so as to identify those who do and those who do not understand.

Such a study was actually done by Jecck, Maccoby, Breitrose, & Rose (1964). The results were surprising. Teachers did not do much better than chance. Moreover, even after some training in interpreting nonverbal cues, they were not very proficient at interpreting classroom cues about students' understanding. Confirmation comes from another study of students' states of mind during instruction. In that research it was found 'that it is virtually impossible to tell what students are thinking from their classroom behavior' (Communication Quarterly, 1984, p. 3).

Unless students are giving very obvious signs of not understanding (lots of puzzled looks and frowns) or very obvious signs of understanding (shaking their heads positively and enthusiastically making notations) teachers should beware of interpreting nonverbal behavior. It appears that the most accurate checking of understanding occurs when teachers ask students to make some sort of verbal response. Questions are the time-honored way to do this. Usually the questions are from the teacher to the student, but students' questions serve also as indicators of students' understanding.

Classroom Questions by Teachers. Although teachers rely heavily on classroom questions to gauge their students' understanding, those same questions may also serve to enhance understanding. The use of classroom questions during instruction may have the effect of providing the following:

Review. Questions asked during an explanation, lecture or other instructional episode may require learners to review the information recently received by them. The students' covert mental processes after a question may be like a "scan" or a search of all the new information. Thus the use of questions may facilitate learning while providing information to teachers about the efficiency of the instruction.

Attention. Questions requiring student responses may heighten student attention during the presentation of material to be learned. Lack of attention results in lack of learning.

Self-awareness. When a question makes students aware that they have not comprehended a point, they may show increased motivation to learn subsequent sections of the presentation. When teachers follow up incorrect responses with correct information, the questions and answers are seen to have direct instructional effects.

Practice. A response to a question allows students to practice stating their recently acquired understanding. Opportunities to practice usually increase the amount learned. Practice also slows down forgetting.

Emphasis. Key issues in the presentation of information can be given emphasis by means of the teacher's questions. This provides students with signals about what is important to know.

Apparently, both students and teachers can benefit from the use of classroom questions. Research from many studies supports this belief. In general, there is a positive correlation between the frequency of classroom questions and student achievement in those classrooms (Dunkin & Biddle, 1974). Perhaps that is the reason why the rate of classroom questioning is often very high. From studies reviewed by Gall (1970), we estimate the average frequency of classroom questions to be about 150 per hour in elementary school science and social studies lessons, and presumably, several hundred per day for almost all teachers. A study of teachers' estimates of their own rates of questioning proved interesting. It was found that teachers generally misestimate such rates. The teachers estimated they asked about 15 questions per half-hour lesson, whereas the actual count was 42. They also estimated their students' questions to be about eight per lesson, whereas the actual count was one! These rates are equivalent to more than one question per minute by teachers and about one question per month by pupils (Susskind, 1969).

The cognitive level of the questions asked in classrooms (and in workbooks or on tests as well) is a major research issue. To check for understanding, a teacher must frame a question that taps something more than simple memory. Many systems of classifying classroom workbook and test questions have been proposed. The most frequently used system is called Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956), which is used to classify questions by the kinds of cognitive processes they require of students. The levels of the taxonomy are given here in...
Increasing order of the sophistication and complexity of the cognitive processes that are believed to be involved in answering a question.

Knowledge. The ability to recall, remember, or recognize ideas or facts. Example: “Name the last five presidents of the United States.” There is a plethora of useful things to be memorized—poems, dates, mathematical facts and formulas, spelling rules and exceptions, etc. But true understanding involves cognitive processes requiring more than simple memory.

Comprehension. Making use of what is received in instruction, without necessarily relating it to other things or seeing implications. Examples: “Explain in your own words the reasons for enmity between the Russian and American governments.” “Give some examples of protein-rich foods.” Some researchers believe that even success at this level is not yet true understanding.

Application. The ability to use abstractions, rules, principles, ideas, and methods in appropriate situations. Examples: “Use the barometer to predict weather.” “Determine a healthy diet.” All experts agree that this ability demonstrates understanding.

Analysis. The ability to break down communication into its constituent parts or elements. Examples: “Determine the different parts of a news story.” “Compare and contrast the capitalist and communist economic systems.”

Synthesis. The ability to take pieces, elements, or parts of things and recombine them into a new pattern or structure. Examples: “Prepare a plan for student government.” “Determine the lessons learned from the history of the Vietnam War.”

Evaluation. The ability to decide if criteria have been satisfied. Examples: “Do you like this play?” “Do welfare programs work?” “Is violence ever justified?”

When records of the teacher's classroom questions, workbook questions, and test questions have been catagorized using this taxonomy, it has been found that very large percentages, often over 80%, of all such questions were at the knowledge level (Gall, 1970). Knowledge-level questions do check the contents of students' memories, but that is not the same thing as checking whether students have understood the instructional material. We note also that taxonomic classification depends on the students' previous experience. If students have previously heard or read the answer, the question requires mere recall, no matter how complex the question may seem. Still, getting teachers to frame more complex questions appears to be an enduring educational problem.

How can teachers break the habit of asking so many lower-order (knowledge-level) questions? The answer is practice, practice, practice. Studies of teachers learning to ask other than knowledge-level questions have shown that teachers can learn such skills. In just a few hours of instruction, including detailed analysis of their classroom performance in asking questions, teachers changed their behavior (Borg, Kelly, Langer, & Gall, 1970). The changes in the question-asking behavior of teachers who received this kind of training have proved to be long standing, with many positive effects still discernable three years after training was completed (Borg, 1972). Thus, classroom questioning at a level of the taxonomy high enough to check understanding is a learnable skill. It is not, however, commonly observed in classrooms.

Concern about the level of questions asked by teachers is relevant only if there is some relationship between questions at different levels of the taxonomy and achievement. Lower-order questions certainly have a place. All children, but particularly the young and the low-achieving children of the lower social classes, need a high level of success in their instructional activities (Brophy & Evertson, 1974). Lower-order questions can provide that success. The frequency of such questions was found to correlate positively with achievement for young, lower-achieving students. But higher-order questions, particularly with middle school and older children, seem also to be important. It is estimated, for example, that if an eighth-grade social studies student scored at the 50th percentile on an achievement test with a teacher who asked predominantly lower-order questions, that student would score at the 77th percentile on the achievement test taught by a teacher who asked predominantly higher-order questions (Redfield & Rousseau, 1981). In addition, Dillon (1982) points out that teachers who ask relatively more higher-cognitive-level questions also tend to elicit student behavior at relatively higher levels. Although the match between the cognitive level of the student's answer and the cognitive level demanded by the form of the teacher's question is high (the odds are only 50-50 that they will match), higher-order questions still manage to elicit student answers that are more complex than do lower-order questions. Asking higher-order questions "works" in the sense of making questions respond to teachers at higher levels.

Teachers who use classroom questions, particularly higher-order questions, to check understanding should keep in mind that analyzing, synthesizing, evaluating, and even just plain remembering take time. Teachers have often been found not to wait long enough after a question to ensure that students have processed their answers. This interval before a student responds to a question is called wait time. (Waiting after a student answers and waiting before a teacher responds to a student's answer are also aspects of wait time.) Teachers generally wait less than one second before requesting a student response. If that wait time is increased to an average of three seconds, through relatively simple training procedures, ten beneficial effects occur (Rowe, 1974):

1. The length of student response increases.
2. The number of unsolicited but appropriate responses increases.
3. Failures to respond by students decrease.
4. Confidence, as reflected in decrease of inflected (questionlike tone of voice) responses, increases.
5. The incidence of speculative responses increases.
7. The incidence of student-to-student comparisons of data increases.
8. The frequency of student questions increases.
9. The incidence of responses from students rated as slow by teachers shows an increase.
10. The variety in types of actions taken by students increases.

Teachers need specifically to provide enough wait time for the students they regard as less able. Teachers have often been found to give less time to these students than to students whom they believe to be more able (Rowe, 1974; Good, 1983). The data supporting the many beneficial effects reported for increased wait time are very impressive. For example, Fagan, Hassler, and Szabo (1981) worked with 20 ele-

BUT DO THEY UNDERSTAND? 269
mentary school teachers of language arts, who led discussions about literature. The teachers received training in wait time, in higher-order questions, in both, or in neither. Student responses were of greatest length for teachers who were trained in wait time. The students of the teachers who were trained to ask higher-order classroom questions gave more alternative explanations and a greater number of higher-order responses. In the classes of teachers who were trained in both wait time and higher-order questioning, there were a reduction in the number of teacher questions, an increase in the cognitive level of teacher questions, an increase in the cognitive level of student responses, and more alternative student responses. Tobin and Capie (1982) found something similar. Wait time of three seconds together with relevant, clear, higher-cognitive questions was found to improve student attention and achievement in science in grades 6, 7, and 8. They found that wait time correlated substantially with achievement (r = .69). Beneficial effects are shown in dozens of studies (Tobin, 1983), and the technique appears to be particularly successful in fostering more complex language from minority groups, such as American Indians (Winterton, 1976).

Now suppose that you waited the requisite three seconds and received an unacceptable response. What should you do? Two alternatives are probing (asking the same student additional questions) and redirecting (asking the same question to a different student). Since both of these teaching techniques are likely to maintain or increase student involvement in the lesson, they are desirable teaching techniques. They also provide the teacher with additional information about the degree to which students have or have not understood instructions. In addition, there is some slight amount of evidence that they also foster learning. In a study by Wright and Nuttall (1970), the frequency of the teacher's redirection of nonanswered questions correlated positively (r = .54) with student achievement. The sample was 12 third-grade teachers who taught lessons on nature study. Brophy and Evertson (1974) found redirection to correlate positively with achievement for middle-class students but suggest that probing is a better teaching technique with lower-class students. Probing can then be used to reduce or eliminate the high frequency of "I don't know" responses among such students. This kind of probing can take the form of rephrasing the question, giving clues, and reminding the student of what's already been established, among other techniques. The frequency of such probing procedures correlated positively with achievement (Brophy & Evertson, 1974; Anderson, Evertson, & Brophy, 1979). In addition, probing to improve and clarify answers that have been given (as opposed to probing simply to elicit an answer from a student) was also correlated with achievement (Soar, 1966; Spalding, 1965). This kind of probing helps students to clarify their answers, to develop generalizations, and to hypothesize outcomes and solutions to problems. Such probing needs to be gentle and nonthreatening, as well as precise, to elicit or improve answers to teacher questions. These probes inform the teacher about the depth of a student's understanding as students are led to elaborate on their cognitive processes.

We must remember that questions have instructional effects as well as usefulness for checking understanding. The questions should be answerable questions, because student success is an important factor to consider in instruction. Furthermore, the questions should often be of a higher order; too many questions are not. Teachers need to monitor student answers to their questions because often the answers do not match the cognitive level that the questions were designed to elicit. Student answers can be responded to by redirection and probing, teaching techniques that help determine the breadth and depth of understanding in a class. Finally, we should remember that there are students for whom classroom questions can prove embarrassing. American Indian students, for example, may be socialized not to answer or ask direct questions in order to avoid standing out among their group. For them, the very shy, the speech- or hearing-impaired, and other subpopulations, questioning as a means to infer understanding may have to be more private (more will be said on this point when we discuss conferencing).

Classroom Questions by Students. The reason the rate of student question asking in classrooms is so low is simple. Most teachers discourage real questions from students. It is often costly for a student to ask for another explanation or to express genuine perplexity, yet that is precisely what is needed by teachers who care about checking for understanding. Teachers have to learn to provide a positive climate for student questions to be asked. This means asking for questions and actually waiting long enough for them to be generated, rewarding students who ask questions, appointing student questioners for the day or week, and otherwise encouraging inquiry. When teachers create such environments, students ask interesting questions, and these questions often promote more student-to-student interchanges. Student answers to student questions tend to be more complex and longer than are student answers to teacher questions (Dillon, 1983). Furthermore, just as the level of cognition that is reflected in a teacher's question can be raised by training, so too can the cognitive level of students' questions be raised (Glover & Zimmer, 1982). In one month, using teachers' praise as reinforcement, fifth-grade children showed 20% to 40% increases in asking questions at the application, analysis, and synthesis levels.

Students' questions inform teachers directly about student understanding of what has been taught. Despite the obvious benefits of having instruction take place in a way that allows students to express their misunderstandings genuinely and to ask for clarification and elaboration, classrooms with high rates of student-initiated questions are rarely observed.

Conferencing. One of the best ways for teachers to find out what students do and do not understand is to take them aside and talk to them in private. This technique is not often observed in practice, perhaps because it is so time-consuming. When it is done, however, conferencing becomes more like ordinary conversation than a question-and-answer session. Teachers can use more declarative statements to elicit responses from students. Though not generally recognized, declarative statements ("That's an interesting idea"); (I've often thought that myself") often evoke longer and more complex student responses than questions (Dillon, 1982). In a private conference, teachers can invite some students to elaborate on their thoughts in ways that are not possible in classroom settings. The conference is particularly useful for learning what is known by students who do not participate in classroom activities. Many experts believe that nonparticipants in classroom activities and discussions should not be "forced" by direct questioning into attempting responses. There are five reasons for this (Maier, 1963): (1) Quiet students may have nothing to contribute at that time; (2) the questions may be perceived as threatening in some way, as they might be to minority, bilingual, speech-handicapped, or shy students; (3) other students might wonder why one individual gets special attention; (4) the questioner's actions
Semiformal Methods

Semiformal methods of checking student understanding are not any more or less demanding than informal methods. Both require considerable skill and thoughtfulness. The semiformal methods do, however, require more teacher time for preparing materials and evaluating student products. In the informal methods, the rapid verbal interchanges require immediate analysis of student responses and immediate decisions about how to proceed. In the semiformal methods, the records of students’ responses are more durable, allowing the teacher the luxury of more time to analyze the student responses and to decide what to do next. The two semiformal methods we will discuss are the assignment and monitoring of practice in classrooms and the assignment and analysis of homework.

Classroom Practice. Students spend a good deal of classroom time working individually on workbook exercises, ditto sheets, or practice problems and assignments given out by the teacher. In the elementary school, as much as 60% of the school day may be made up of such seatwork assignments. In secondary schools, the percentage is lower but still significant. However, when questioned, students doing seatwork often do not know what they are doing or why; they do the assignments merely to finish them and get another assignment (Anderson, 1985). As noted earlier, teachers must be explicit about what their students are to do and how to do it. From Anderson’s research on students’ cognitions during seatwork, we see that teachers also need to tell students why they are doing something. Students who mindlessly do practice problems in workbooks and on dittos have stopped trying to understand anything. Schooling does not make sense to them. Finishing tasks is all they know. To guard against this apparently widespread problem, teachers need to remind students regularly about the relationship between what they are doing and the instructional goals for the class.

The best practice problems offered in workbooks, on dittos, or for boardwork are those that provide the teacher with diagnostic information about students’ understanding. Carefully constructed diagnostic practice problems can reveal student thinking processes, particularly when students give wrong answers. It is easier to see how a diagnostic problem works by using an example from mathematics, but the principle is applicable to all subject-matter areas. For example, let us examine the simple problem 248 - 59 = _. You might get the answer 189, which is correct and provides you with no further information. But if you have planned your practice problems with diagnosis in mind, the incorrect answers 211, 299, and 99 can provide you with diagnostic information of very high value. This is because these answers are perfectly correct for people with faulty methods of subtraction. If the method used is always to subtract the smaller number from the larger number in a column, no matter which is on top and which is on bottom, the “right” answer for a student with that fault in understanding subtraction is 211. The student has an algorithm to solve the problem, but it is wrong. Using the language of computer programmers, this is called a “buggy algorithm,” as when computer programs have “bugs” in them (Brown & Burton, 1979). An answer of 211 to our sample problem allows the teacher to hypothesize the nature of the bug in the student’s subtraction program. With accurate diagnosis, remediation is much easier. Other answers to this simple problem indicate other buggy algorithms a student might be using. For example, the answer 299 indicates that the student is correctly borrowing from the next column over, as needed, but forgets to reduce the numeral in that column by one. Thus 299 is a perfectly correct answer when the student applies a consistent and systematic but completely wrong method of subtraction. The answer 99 indicates another buggy algorithm; it is correct if the algorithm being used is to borrow, when necessary, from the leftmost column of the upper numbers instead of borrowing from the very next column to the left.

The point of this discussion is that student errors are more informative to teachers than correct responses. Errors that are neither random nor careless allow teachers to diagnose problems in students’ understanding. Questions to elicit such systematic errors of thinking are difficult to construct, particularly in content areas other than mathematics. Nevertheless, an odd answer to a social studies question might indicate a student’s lack of understanding of the differences between communism and socialism. Without such definitional issues clearly in mind, certain kinds of interpretations of social studies problems will necessarily be inadequate. In science, a student’s inadequate answer may indicate a failure to understand a principle such as the relationship between force and mass. In literature, a student may interpret the motives of characters based on current conceptions of the roles of males and females instead of those prevalent in Chaucer’s time. In each of these cases, students’ incorrect statements or answers indicate bugs in the problem-solving algorithms that they are using. As often as possible, the teacher’s goal should be to use the practice problems in workbooks, on dittos, or on the board to elicit any buggy algorithms that are present. The development of such diagnostic questions is difficult; nevertheless, for hard topics and difficult concepts, a teacher can build up a good set of diagnostic questions over time.

When incorrect answers are due to carelessness, feedback can help students. When students provide no answers or give random answers to practice problems, reteaching is probably required. When students make consistent and systematic but wrong responses, debugging of the students’ logic is necessary.

Some other points about classroom practice problems of all types may be helpful. Deep and enduring understanding of subject-matter areas and special topics comes from mastery of basics. It is not easy to progress in physics without thorough understanding of some basic principles, usually learned by extensive practice over a wide range of problems. Math, bridge, chess, historical analysis, curriculum analysis, oil painting, and literary criticism all require practice with the basics in each field. Therefore, practice can be defended as necessary for the development of understanding. Recent research in mathematics instruction (Good, Grouws, & Ebmeier, 1983) identifies practice opportunities as a key element in effective instruction. And Ross-White (1985) also identifies practice as one of the important instructional functions that every teacher needs in order to create successful instruction.
There is also evidence that overlearning material, through repetitive practice, affects what is learned as well as how much can be retained over long periods of time. For example, when students overlearned unfamiliar scientific information, they were found to transform the material into personally meaningful ideas. They were better able to figure out the main conceptual idea in the material they had learned. The students' repetitive practice was apparently the basis for an increase in problem-solving skill and in transfer (Mayer, 1983). In other studies of reading and mathematics (Samuels, 1985; Resnick & Ford, 1981), when certain levels of automaticity were reached by means of practice, students were found to go on to comprehend more or inven new ways to do the problems they had to master. So practice is not just to provide an opportunity to check understanding or to ensure retention. It is usually a necessary step in the development of deeper forms of understanding.

**Homework.** Homework is often another form of practice. Thus what has been said about classroom practice is also applicable to homework. In particular, home assignments of practice should be doable, providing students with opportunities for success. And the homework should also be diagnostic, giving teachers information about the ways in which students are thinking about the content they must learn. In addition, when assigning homework, teachers must be concerned about equity, feedback, and ties between the homework and the curriculum.

Equity issues become important because students' homework completion rates are not always random. They may be associated with racial, ethnic, or socioeconomic factors. It is possible that heavy reliance on homework assignments in classes of mixed social status provides the middle-class students with an advantage on school achievement measures due to differences in family interests and monitoring of schoolwork. Teachers need to be aware that the monitoring and help with homework that is characteristic of some families is not characteristic of all. In some programs of school improvement, parents have been enlisted to help monitor homework to ensure that it gets done. But not every family has members with the time or knowledge to help a young student with homework. Furthermore, in some families, chores take precedence over homework. Teachers who assign homework to provide students with practice and themselves with a sample of student behavior from which to analyze student thought processes must also consider the variability in family support of homework in their classes.

Feedback is necessary for improved performance in almost all learning situations. If students turn in homework, it deserves to be evaluated. Teachers who assign homework to diagnose problems or provide practice but who do not regularly evaluate student products not only miss the opportunity to help improve student learning but also foster students' belief in the controlling nature of schooling through the assignment of mere busywork.

Homework is a way to control student time at home. Therefore, to take that time away from students, teachers should be sure that the homework assignments have clear instructional purposes. Students are quick to perceive busywork. Their compliance with homework assignments should not be confused with their real attitudes toward schooling, which become negative if they feel too controlled, particularly at the junior and senior high school levels. The easiest forms of homework assignments to defend are those tied closely to the curriculum of the classroom. The classroom curriculum, of course, needs to be aligned with the goals and outcome measures of a district. When there is congruence among all three (goals and outcome measures, classroom activities, and homework), student learning is enhanced.

Homework is not just an opportunity for teachers to evaluate student products. Research indicates that there are direct instructional effects associated with the assignment of homework. Correlational evidence (Wolf, 1979; Keith, 1982) and experimental evidence (Marshall, 1982) suggest a causal relationship between the number of homework hours per week and achievement. The most defensible statement to make about homework is that if assigned in some sensible quantity, if tied to the curriculum, if success is possible or diagnostic information can be derived, and if feedback is regularly provided, homework will be of great utility to teachers and students in fostering learning.

**Formal Methods.**

The most frequently used method of assessing students' understanding of what they have been taught is the test. A heavy reliance on tests—in particular, multiple-choice tests—is a characteristic of the United States of America. Other countries are not nearly so test-oriented, nor are they as enamored of the multiple-choice test. In fact, in many countries a student can go from elementary school to graduate school without ever having encountered a multiple-choice question!

Tests designed to assess students' understanding of what has been taught in schools are usually called achievement tests. A teacher who is informed about achievement testing is aware of at least three things: (1) the purposes and types of achievement tests that exist, (2) the types of items used in achievement testing, and (3) the technical characteristics of acceptable achievement tests.

**Purposes and Types of Achievement Tests.** Let us first distinguish between two major types of achievement tests to assess understanding: standardized achievement tests and teacher-made tests of achievement. Standardized achievement tests are usually designed by states or commercial publishers to assess a wide range of school development. The information obtained from such tests is for reporting to parents, local school boards, and state and federal agencies about how well or poorly schools, districts, and states are doing in the preparation of students. The tests often do a poor job of providing teachers with useful information about what individual students understand, and they rarely provide diagnostic information to modify instruction. The usual practice is for standardized achievement tests, such as the Metropolitan Achievement Test (MAT), the Stanford Achievement Test (SAT), and the Iowa Tests of Basic Skills (ITBS), to be the cornerstone of a summative evaluation. The tests measure what has been learned after instruction has taken place. Moreover, such tests measure the products of that instruction and provide no information about the processes used by students as they attempt answers to the test items. Finally, standardized achievement tests often measure memory or the comprehension of information and only rarely assess higher and more valued kinds of understandings.

There are, to be sure, standardized tests that claim to diagnose learning disabilities, neurological handicaps, or reading problems or claim that they can identify gifted, retarded, or achievement-oriented students so that instruction can be modified to accommodate their exceptionality. But these standardized tests are usually achievement tests; they are aptitude, ability, intelligence, interest, or personality tests.
Nor do standardized achievement tests provide teachers with information that can guide the course of classroom instruction. Thus we will focus on the purposes and types of teacher-made tests.

A teacher's testing for understanding will be guided by concerns for obtaining information that allows for the placement of students, revision of instruction, diagnosis of student problems, or the certification of competency. When information for the placement of students is desired, as when a new student enters the class, a wide-ranging set of questions of different degrees of difficulty and tapping different levels of cognitive functioning is wanted. From such a broad sampling of student cognition, a teacher may decide where in the curriculum or with what instructional groups the new student should be placed. Sometimes, minimal competency testing is called for, as when a teacher must decide if a student has enough algebra to go into the trigonometry class or if a student understands enough about the mechanics of writing to skip the technical writing class and go directly into the creative writing class. Such competency tests for placement are usually short, with relatively easy questions. Their purpose is to determine if certain basic skills are present.

Teachers who want information for revision of instruction create formative tests—tests that give both teachers and students information on how they are doing as instructors and as learners. Such tests take such forms as the 10-item mathematics facts test; a short-answer test requiring definitions of the concepts of ecological niche, adaptation, and natural selection; or a request for a one-page description of how a bill gets through Congress. Formative tests are usually short and are given frequently, perhaps every week or two. They are tied to instructional objectives or to the domains of knowledge that must be mastered. They are often graded in class, or overnight, to provide rapid feedback to teachers and students so that they can modify the process of instruction, if necessary. Such tests are often given back to the students to review or to keep so that they may learn from any errors they might have made. Formative tests permit teachers and students to monitor their progress.

Diagnostic tests are perhaps the most difficult to build. The goal of the diagnostic test is also to provide information that is formative, to revise instruction, but diagnostic tests go beyond measuring progress through the curriculum. They are built to inquire into the cognitive processes used to solve the problems presented to the students. Diagnostic tests must reveal the students' thinking. They concentrate on uncovering the cognitive processes elicited by instruction. Such tests are less oriented to the products of instruction. Diagnostic tests can be quite short, since diagnostic items are so difficult to create.

When it is necessary to attest to the fact that a student has learned something, without any great concerns for placement, formative information, or diagnosis, a summative test is appropriate. Such tests are given at the end of an instructional sequence and are used either to assign grades or to certify some degree of mastery of a body of knowledge. The products, not the processes, of instruction are attended to in scoring such a test. The interpretation of summative tests, however, is a very complex problem.

There are two major ways of interpreting summative tests (and, for that matter, formative, diagnostic, and placement tests as well). The first is with regard to a criterion, and thus such tests are called criterion-referenced tests (CRTs). The second is with regard to norms, and thus such tests are called norm-referenced tests (NRTs). A CRT measures whether a student has or has not reached a criterion or some particular specified level of achievement. A test score depends on the specification of an absolute standard of quality. This standard is independent of a student's actual score. If a student reaches or exceeds the standard, he or she will be judged to be a master, proficient, minimally qualified, able, or whatever. Everyone who passes such a test is proficient, and everyone who fails such a test is not. These tests are constructed in the same way as tests to obtain a driving license. The most desirable situations for using CRTs are for measurement of achievement in individualized programs of instruction, where skills and content areas are often well defined and where students work on different curriculum units. The CRTs do an excellent job of estimating student ability in a well-circumscribed domain, such as two-column addition with regrouping, foreign language verb conjugations, or determining molecular structures. The CRTs serve well for monitoring progress through the curriculum, a formative use, and can be created to diagnose problems as well. Most important, for summative purposes, a CRT measures what a student has actually learned about some domain of knowledge.

On the other hand, with an NRT a student's performance is interpreted in light of the scores achieved by others taking the same test. The NRTs are very useful for classifying students (A knows more than B, and B knows more than C) and for selecting students when quotas exist (placing the top 15% on a science test into an advanced course in science). Unfortunately, with most NRTs the information from scores and grades does not inform us about what a student actually knows. Teachers make judgments about who knows more and who knows less. An NRT may inform us that some students understand physics better than other students, as evidenced by their higher grades or percentile standing on a midterm test of physics achievement. But even when a student is in the 90th percentile or in the A category in the grading system, the amount of physics the student actually knows is never clearly revealed. Only the student's relative standing in comparison to other members of the group that took the test is known. With NRTs, there are winners (higher scorers) and losers (lower scorers). With CRTs, it is possible for everyone to win (pass) or everyone to lose (fail) because the criterion for success or failure is chosen by knowledgeable people, and that criterion exists independent of the students' performance.

Which kind of test is better for assessing understanding? That depends on your purposes for testing, the decisions you intend to make from the information you obtain, and the nature of the domain of knowledge in which you are assessing understanding. The domain of knowledge is an important consideration because it is easier to construct CRTs in some content areas than in others. For example, subtraction of up to five columns, with regrouping, is an easy domain from which to create CRTs. Furthermore, it makes no sense to say that Sally is better at subtraction than Don; it makes sense only to say whether Sally and Don can or cannot subtract. When the instructional objective is derived from a well-specified domain of knowledge, such as identifying the writing styles of such American writers as Hemingway or Steinbeck or finding solutions to biochemistry problems or learning geography facts about states and nations, a CRT is not difficult to create. Teachers can set the criterion level where they want it to assure that students meet their standards. Whenever possible, the assessment of students' understanding of school subjects, particularly for summative purposes, should be done with CRTs. Unfortunately, some domains of knowledge do not easily lead to CRT development. Moreover, the creation of enough items at higher levels of cognitive processing to make up a useful CRT is often difficult.
Regardless of the type of test developed and the purposes guiding that development, the key to a good test of students' understanding is having good test questions. We turn now to this issue.

Types of Items Used in Achievement Testing. One way to classify item types is by means of the taxonomy for classifying objectives and questions presented earlier, whereby the level of cognitive processes that are expected to be used by a student is inferred. Another way to classify questions is by the type of responding called for by students: recalling, recognizing, writing, selecting, and so on.

As shown in Figure 11.1, the major types of test items are essay questions and short-answer questions. Short-answer questions can be subdivided into those that require a student to supply the answer and those that require the student to select the answer. The select type can be broken down further into true-false, matching, and multiple-choice items. Essay items, as compared to short-answer items, are often heralded because of their potential for calling forth higher levels of cognitive processing, such as analysis and synthesis, which may be required to answer compare-and-contrast-type questions or to answer evaluative questions (see examples in Figure 11.1). While this is generally true, a well-crafted multiple-choice item can also tap these levels of processing. Such an attempt at tapping higher-level processes is seen in the examples of multiple-choice items in Figure 11.1. These are not intended to be simple memory-level items (though, of course, they could be).

Essay questions are also reputed to be easy to prepare in comparison to select-type questions. But good essay questions do not come forth perfect, ready to be answered. The good essay question writer must prepare an answer to the question and design a scoring key for the essay as well. Thus the easy-to-prepare criterion rarely holds for serious essay question writers. The essay question does two clear advantages over short-answer questions: It eliminates the effects of guessing, and it tests writing ability, an important communication skill that is often underemphasized in the schools.

The short-answer formats have some decided advantages, too. They allow a greater sampling of domains of knowledge, thus ensuring that the test covers many curriculum content areas. They are scored more easily and more accurately. Essay tests, with their complex and even creative answers, are more open to interpretation than short-answer questions. Essay answers take longer to score, and errors, misjudgments, and bias in scoring can occur. Nevertheless, essay, even short-essay tests, are the preferred way to get information about a student's knowledge, reasoning, opinions, and creativity.

Of the short-answer questions, the supply type, short-answer recall, and high-grade multiple-choice items allow a teacher to tap into some of the higher levels of cognitive process. True-false, matching, and run-of-the-mill multiple-choice items generally tap memory-level processing only.

What does it take to make a "good" test? Nitko (1983) presents the steps in that process (Figure 11.2). The all too typical classroom test, slapped together in minutes, does not represent the best in the art of achievement test design, as Figure 11.2 makes clear. For more on the art of item writing, see Wesman (1971); for on testing, see Nitko (1983), Cronbach (1984), and Popham (1981).

Technical Characteristics of Achievement Tests. Among the most important decisions to be made by a teacher is whether the items of the test for assessing
understanding are derived from the curriculum and match the goals for the course or instructional unit. There is technology to help a teacher do this. It is called the behavior-content matrix, and is the way to accomplish the task specified in box 6 of Figure 11.2. First, a teacher must set out columns and rows. The columns are labeled with the types of cognitive behavior that are desired from the students. The Bloom taxonomy provides one way to describe student behavior. An abbreviated version of that taxonomy is provided as an example in Table 11.1, but any designations of student behavior will do as well. The content to be covered is listed in the rows. These can be listings of curriculum units, concepts to be learned, or any other systematic description of the content in an instructional unit. The cells of this matrix are our focus. They can be filled in with questions that tap the important behaviors and contents that should be covered in a test. Table 11.1 shows how this is done. A behavior-content matrix ensures that a teacher-made test covers the important areas of a course or curriculum unit. This is true whether a CRT or an NRT is desired. Either kind of test can be developed to cover the areas in a behavior-content cell or to cover all the desired behavioral competencies across a row of important content.

A test may also cover an entire course, and the matrix helps to delineate all the areas needed to develop a comprehensive test. In fact, to ensure fairness in creating a test, the behavior-content matrix can first be filled out in a way that ensures that a teacher includes the items that match his or her judgment about the most important topics or issues. An example of how this is done is given in Table 11.2. After the rows and columns of a behavior-content matrix have been created, but before questions are generated, a teacher can weight the matrix. First the teacher can examine the rows and give more weight to content of greater value and emphasis. Then the teacher can weight the columns to emphasize the student behaviors of greater value and emphasis. The row and column percentages in Table 11.2 reflect this kind of weighting for an instructional unit on American novels. In each cell we see the approximate number of questions for a 30-item test that is designed to reflect the teacher’s values and instructional emphasis when teaching these American novels. This version of a behavior-content matrix is called a table of specifications—it specifies how to construct an achievement test in a rational manner. Such tables of specifications help avoid generating too many items of one type or too few of another type.

One of the most useful things a teacher can do is generate the behavior-content matrix and tests for assessing instruction before instruction begins. The behavior-content matrix and table of specifications can be used to guide the time allocations and content choices that every teacher must make. When this is done, a technical feature of all good tests is automatically taken care of: the content validity of the test. When teachers serve in their role as achievement test developers, the most important of the many kinds of validity they must be concerned about is content validity. Content validity refers to the degree to which a test or test items measure the instructional domains that students have had an opportunity to learn.

Content validity is a logical procedure, dependent on sound judgments about how items relate to the curriculum that was taught. For example, if content is added to a social studies curriculum, say on the People’s Republic of China, and students study that content, a teacher’s test should include the new material. Even more important, if a unit of instruction is dropped, skipped, or no longer stressed, as when American concern for problems in Southeast Asia waned after the Vietnam War, a
test should not still have items about American involvement in Vietnam. Students have a right to demand from teachers and schools the opportunity to learn what is on the achievement tests used to judge their competency. A behavior-content matrix and a table of specifications guide instruction and test-item development to ensure congruence, overlap, or alignment between curriculum and outcome measures. This provides students with the opportunity to learn what is on the test and helps ensure that the test has content validity.

The last important technical consideration for thinking about how teachers can assess understanding by classroom testing is the concept of reliability. For the kind of teacher-made tests of achievement that we have been discussing, reliability refers to the dependability, consistency, or overlap between test results. If one measurement occasion to another. The ruler can be used differently on two occasions according to the angle at which you view the teacher-made tests of achievement that we have been discussing, reliability refers to the dependence, consistency, or overlap between test results. When the reliability of a test is derived, it is possible, but rare, for decisions about whether students should be given or denied entry to certain classes or to programs designed specifically for the gifted or students of low ability. And reliability should be determined when we assign grades that will influence how others will view the student's competency. Determining reliability is not difficult to do, if we are going to make important decisions about our students, we should formally determine the reliability coefficient of the tests we are using. This would hold for decisions about whether students should be given or denied entry to certain classes or to programs designed specifically for the gifted or students of low ability. And reliability should be determined when we assign grades that will influence how others will view the student's competency. Determining reliability is not difficult to do, though it is time-consuming. (The methods for computing reliability can be found in Popham, 1981, and Cronbach, 1984.)

Reliability refers to the dependability, consistency, or stability of the scores or decisions about students that you must make. When the reliability of a test is derived, it is often presented as a number between 0 and 1 (it is possible, but rare, for reliability indices to be in the negative range of 0 to -1). The closer the value to +1, the more reliable the test. The numerical determination of this figure can be made in many different ways, but conceptually it is done as follows: Develop a large pool of items on the topic of interest, split the items randomly into two parallel forms of a test, administer the two forms of the test to a large and heterogeneous group of students, and determine the correlation coefficient between the two tests. If students score high on both forms of the test, we infer that their scores appear to be dependable.

When we measure achievement, our instruments, our tests, are not always very good rulers. Our students also vary considerably in what they can retrieve from their memories on different testing occasions. And our scoring of results, particularly with essay tests, is also subject to variation. The problems of achievement testing are therefore very difficult because our measurement is filled with error. The way we estimate the error in tests is by means of a statistic called the reliability coefficient. If we are going to make important decisions about our students, we should formally determine the reliability coefficient of the tests we are using. This would hold for decisions about whether students should be given or denied entry to certain classes or to programs designed specifically for the gifted or students of low ability. And reliability should be determined when we assign grades that will influence how others will view the student's competency. Determining reliability is not difficult to do, though it is time-consuming. (The methods for computing reliability can be found in Popham, 1981, and Cronbach, 1984.)

Reliability refers to the dependability, consistency, or stability of the scores or decisions about students that you must make. When the reliability of a test is derived, it is often presented as a number between 0 and 1 (it is possible, but rare, for reliability indices to be in the negative range of 0 to -1). The closer the value to +1, the more reliable the test. The numerical determination of this figure can be made in many different ways, but conceptually it is done as follows: Develop a large pool of items on the topic of interest, split the items randomly into two parallel forms of a test, administer the two forms of the test to a large and heterogeneous group of students, and determine the correlation coefficient between the two tests. If students score high on both forms of the test, we infer that their scores appear to be dependable.
The same is true if students are at about the middle in scores on one form and also at about the middle in scores on the second form. And you would expect low scorers also to show such consistency over the two test occasions. Such a test would produce a high correlation between the forms, indicating stability, dependability, and a sense of assurance.

Now let us consider the opposite case: Suppose students were found to have very different scores from one test occasion to another. The correlation coefficient would be low, indicating instability, a lack of dependability, and little assurance that the scores on the test are trustworthy. Measurement of height from one occasion to another sometimes may yield correlation coefficients of about .99 or .98. Measuring achievement with the same kind of tests from one occasion to another under the most optimal conditions yields correlation coefficients of about .92 or .90.

So it appears to be possible to build highly reliable tests from which we can make important decisions about our students. With such tests we can feel that our decisions are being made on the basis of dependable data. But in most classrooms the tests built by teachers rarely come close to these levels of reliability. Two forms of a test to assess reliability are not usually created, and thus we must find ways to estimate reliability from a single administration of a test. Also, the typical classroom test is not very long, so we do not get a chance to measure the full range of a student's achievements (it's like estimating height from measuring the length of a person's leg). This technique works, to some extent, but it is not nearly as good as measuring the entire body. Furthermore, the test items are not carefully screened and tried out to get rid of ambiguities, bias, and excessively hard or easy items that provide no useful information. Finally, the size of the student group is small, and in a single class the full range of abilities is usually not represented. All these factors come together so that the typical teacher-made test of classroom achievement has a reliability in the range of .50 to .70. This magnitude of reliability is just barely good enough to obtain useful information to guide instruction, but it is not good enough to make important decisions.

There is a direct relationship between the reliability of a test and the magnitude of the error a teacher may be faced with when interpreting a student's score. Whenever we have less than perfect reliability, which is just about all the time, we need to think of scores as bands, not as precise points along some scale. That is, the lower the reliability of a test, the greater the band in which a person's score can be truly located. A person with a score of 68 on a test with perfect reliability (+1) is acknowledged to have a true score of 68—there is no error to worry about. In a test with a reliability of .90, that same person may (through statistical techniques) be judged to have an error in measurement of 2, indicating that if we wanted to be reasonably sure of that true score, we would have to think of it as somewhere between 66 and 70 (±2 points). In fact, to be very certain, we might estimate that the true score was between 64 and 72 (±4 points). With anything less than perfect reliability, we are not justified to think of a score as a precise point. In the reliability ranges of many teacher-made tests, around .60, the error might easily be about 8 points, indicating that the true score of a person who received a test score of 68 is probably between 60 and 76 (±8 points) and almost certainly between 52 and 84 (±16 points). That is a very large range indeed! And that is why you should determine reliability—it tells you how much error there is in your scores. You must have that information in order to judge if you can trust the decisions that will be made on the basis of the test scores.

When dealing with criterion-referenced tests, the computation of reliability is somewhat different. Since only a few items may make up a CRT, we look for the dependability of the decisions that must be made. For example, let us assume that a few short-answer items are developed to assess whether or not students understand the reasons for increased urbanization. We might declare that any student getting three or four right out of the four items presented has mastered the curriculum unit on urbanization. Anyone who answers two, one, or zero items correct has not mastered the curriculum. This short test can be administered a second time, and we can look at the percent of students identified as masters the first time and ask what percent of them were classified the same way the second time. These estimates are usually expressed as percentages or probabilities. A probability of .86 indicates that a test is usually correct in identifying people who do or do not reach criterion standards. A test with a probability of .54 is much less useful—it is accurate only half the time. Whether reliability estimates are computed by correlation or by means of estimates based on percentage or probability does not matter. What we learn to do when we think about reliability is to estimate the degree of dependability of our decisions. When we judge our students' understanding through formal tests (or in any less formal way), we must be sure that the methods we use for making decisions are reliable enough for the purposes at hand. The more important the decisions to be made, the higher the reliability of our methods must be.

We have noted that tests can be designed for many purposes but that teachers really need to be most concerned about their own classroom tests of achievement, both NRTs and CRTs. Teachers should be concerned about developing diagnostic tests and must pick item types to reflect their purposes. Test development, if done well, is a laborious process; it requires much time and many opportunities to administer and redesign the test. The use of behavior-content matrices can help teachers assess understanding more rationally and virtually ensures high content validity—one of the crucial elements of a test. Achievement tests that are reliable and have valid content are useful for decisions about students. Conversely, if a teacher's tests cannot demonstrate content validity and reliability, they cannot be used for decisions of any kind. Testing, the most formal procedure for assessing understanding, is also the procedure with the most highly developed technical methods for evaluating the adequacy of the assessment instruments. Thus with sufficient concern, time, and thought, testing can be done competently and in such a way as to be useful to both instructors and students.

QUESTIONS FROM NEW TEACHERS

Q. You gave no example of an advance organizer. Can you give us one?
A. A good example of an advance organizer is hard to come by. Still, in study after study, something that an investigator called an advance organizer worked better than no such thing. Let's try this. Suppose we are to begin teaching a unit on ecology to fifth graders in the northeastern United States. We might start by defining ecology and then say, "We are going to study ecosystems, systems of relationships. Keep in mind what you have heard about acid rain, because acid rain is a good example of how some factor far away can influence what goes on close to home. We will look at systems in balance and systems disturbed by factors like acid rain. We will look at cycles like the rain cycle and chains of interrelated events like the food chain. We will talk of change in nature, how slow it is to get things back to balance once they are
observation instrument must be analyzed. The goal is reflection on what was taught
yourself. The other is to have a colleague observe your teaching. Both provide
we do. Feedback must, however, be data-based. The videotape or a colleague's
concerns need not slow learning down; on the contrary, they can speed it up!

Q. When you emphasize the familiar in teaching children, don’t you slow them
down too much? You can’t help them learn new things if you keep tying them to
things they already know.

A. What you say sounds logical, and certainly you want to worry about the
pace of instruction; but what I am emphasizing need not slow learning or make it
trivial. By making the new familiar, we may be able to speed up learning. So I do not
agree with you. Nor do a lot of psychological theories and data. The British
associationists of the sixteenth and seventeenth centuries discussed how new ideas
and sensory input could only be made comprehensible by means of their relationships
to familiar ideas and impressions. The great psychologist of the nineteenth century,
Herbart, postulated that all new information was understood if it was tied to the
“aperceptual mass”—the storage facility for the familiar. The Herbartians, pedologi-
cal theorists around the turn of this century, taught lesson plan designs to reflect these
ideas, and their techniques still appear to be sound. They emphasized, as did John
Dewey, the role of familiar examples for helping children learn new things. Let me
give you an example from a psychological experiment of a few years ago. Listen to
this passage and tell me what it is about.

He begged and pleaded for years, finally getting his way, thanks to her faith and
her cleverness in obtaining the money. All the people around him thought he was
crazy as he went off with his three sisters. He beat and ran with nature at his
back but never reached his goal, stymied by blocks he never imagined, dying a
failure in his own eyes, missing his chance for fortune.

Does it make a lot of sense? What if I simply say one word: “Columbus.” Instantly
the paragraph is perfectly comprehensible. We know that the three sisters are Niña,
Pinta, and Santa Maria, that the crew thought him mad, that beating and running are
sailors’ terms, that the block was a continent, and that the failure was in not finding
China. One familiar cue, and suddenly the new is totally comprehensible. The un-
familiar, odd, and divergent is rendered understandable. That is why concern for a
learner’s schema and an emphasis on examples is important for a teacher. Such
concerns need not slow learning down; on the contrary, they can speed it up!

Q. OK, but how do you suggest we learn these techniques to explain better?

A. Two ways come quickly to mind. One way is to videotape (or audiotape)
yourself. The other is to have a colleague observe your teaching. Both provide
feedback to you, and it is only on the basis of feedback that we might change what
we do. Feedback must, however, be data-based. The videotape or a colleague’s
observation instrument must be analyzed. The goal is reflection on what was taught
and how it could be taught better. The teaching tips, the model of explaining, and the
questioning techniques have all been taught to teachers in the past. Therefore, it
should be reassuring to know that it has already been demonstrated that such skills of
teaching are learnable and not just the conjecture of researchers.

To me, learning these kinds of skills for teaching is not much different from
learning to be a gourmet cook instead of a slapdash heater-upper. To be a gourmet
cook requires time, practice opportunities, the comments of observers, practice,
creativity, and a sense of balance as well as knowledge of the basic recipe. There are no
real substitutes for thinking, practice, and feedback; if you want to explain better or be
a different kind of cook, those things cannot be avoided. I believe that teachers get
tons of practice and feedback. If we want to improve, they must
have more feedback.

Q. Student questions really are rare. I keep asking, “Any questions?” but the
students simply don’t reply. What can I do?

A. First we have to distinguish between asking for questions as a social conven-
tion and as a real invitation. As a social convention, it is often used by instructors the
way many people use “How are you?” You tend to be surprised when you get a real
answer to that question, like “I’ve had this pain in my back that began three years
ago, so I went to the doctor and...” “How are you?” is really not intended to elicit
information. Neither are most instructors’ requests for questions.

To promote student questions requires training in question asking. You could,
for example, learn to use silences—very long silences—after asking questions. It is
so uncomfortable for students to be stared at by an expectant teacher that they
always break the silence first and will ask a question if you have asked for one. It is
not easy to do, but it works, even with generally uncommunicative students.

Once you elicit a question, however, you are obliged to reinforce the student who
asked. You not only owe the student a thoughtful answer, but you must also say such
things as “Good question!” “That’s interesting!” or “Let’s really examine that one.”
Reinforcement, when appropriate and genuine, increases the frequency of student
questioning.

Another way to elicit questions is to give stars, awards, tokens, or even raises
for good student questions. The tangible rewards make it fun for the students to get
started asking questions. The rewards can be phased out when student question
asking becomes more regular.

It is possible to appoint different students to be official question asker for
each lesson, day, or week. When you are through with an explanation, you can turn
to the designated student and ask for questions. Ideally, the questions would consti-
tute clues about what to reexplain to the whole class.

Remember that teachers can increase not only the frequency but also the quality
of student questions.

Q. What did you mean when you said practice leads to other things besides
achievement of the things that are practiced?

A. Teachers and parents sometimes forget how practice of one thing lets other
desired cognitive events come into play. For example, we may overlook the fact that a
child’s information processing system may be quite limited. The average child may
not be able to concentrate on decoding written verbal material and at the same time
comprehend what those words mean. Comprehension for many children occurs only
after extensive practice with decoding words, to the point that the decoding is almost automatic. The quicker the word recognition, the more likely that the meaning of the word can be understood. Thus practice in decoding and word recognition is actually the precursor to comprehension for many children. And in mathematics, after much practice with mathematics facts, some children have become quite inventive. When it is very easy for these children to answer a question like \( 2 + 4 = \) _, these children often start to invent mental games such as \( 2 + (2 + 2) = \), or \( (2 + 2) + 2 = \) _, or \( (2 - 1) + (4 + 1) = \). That is, after extensive practice with a symbol system, they can invent mathematical rules and mathematical games. These are much more desirable outcomes than simply knowing the mathematics facts themselves. So that is why I said practice has more profound effects than just on the skill areas that are practiced. Practice of simple skills may be a basis for the development of more important outcomes.

Q. You make it sound as if the only good homework is practice of things done in class. I use homework to "stretch" kids. Do you find that wrong?

A. Teachers cannot possibly fit into each day all that they might want to teach and all that society wants them to teach. Homework is a perfectly fine way to add to the basic curriculum. Homework might be used in preparation for a future topical event, such as when students are assigned material to interpret an upcoming space mission or to interpret a National Geographic special on television. In the same manner, homework can be used to illuminate a past topical event, such as a worker's strike in the city or a border war between countries. Homework assignments can be enrichment for some children, as when the most able history students in the class get a special project. And homework can be individualized, so that a student who asks a question gets to do research over a few days or weeks in order to answer the question.

But note that unless the achievement-testing program takes into account these kinds of homework assignments, the effects of such homework will never be known. I think such inventive homework assignments are necessary. They have intrinsic merit. I simply remind you that unless the curriculum and the outcome measures are all aligned, there will be no measurable effects of the homework part of the curriculum. Homework is often outside the standard or expected curriculum, and therefore its effects on student achievement are almost always unknown. Homework shows up as a successful mechanism in promoting measured achievement when it is tied directly to the basic curriculum and the basic curriculum is tied to the outcome measures that are used to measure achievement. Do not think I only approve of that kind of homework. I do not. However, that is the kind of homework that is most likely to boost achievement in a direct and demonstrable fashion.

Q. I don't think that I have seen many criterion-referenced tests. How come? Those sound better than norm-referenced tests.

A. Criterion-referenced tests are not intrinsically better—they are only better for some purposes. When the instructional domain is very discrete, such as with multiplication, a CRT is usually the better test. No one really cares if someone is better than someone else in multiplying. Our society simply wants people to be able to multiply! The reason norm-referenced tests have been developed so extensively is tied into the philosophy and social purposes of schooling. For a long time, schools operated as if one of their jobs was to select and promote the most able students and to drop the least able students along the way. Norm-referenced tests gave information that allowed the least and most able students to be identified. Now that we as a nation have committed ourselves to retain and educate all our youngsters to promote social equality, we find we are still using tests that were designed to promote a different social philosophy. When we need to select the most able people from a pool of people, norm-referenced tests have the edge. But I believe that for most teachers, the measurement of classroom achievement can best be accomplished with a criterion-referenced test.

Q. You make testing sound so difficult. Have all the tests I have taken and given been so awful?

A. I hope not! But remember, tests serve so many other functions. They motivate students to do reviews; they provide practice opportunities; they inform students about progress; they inform teachers about who needs further help or whether the class can move on or not. So even when their technical qualities are deficient, tests can still serve very useful purposes. However, if you were ever given a hastily contructed, relatively short achievement test, over a large body of material, I would bet that the technical quality of that test was quite low. And if important decisions about you or your classmates were made on the basis of such a test, they were probably not very accurate.

Q. Inaccuracy in measurement means that scores kind of float around—they can be as much as 5% or 10% higher or lower. How can I ever assign grades with such uncertainty?

A. It is true that reliability for any given test may be low, but the reliability of a set of tests can give you accurate information. In addition, you can use other indicators of competency, such as essays written at home, other home projects, and classroom participation. Recognize that grading is serious business. You are judging others and affecting their life chances. Such judgments should be made on the basis of as much information as possible and be done as humanely as possible. When you start teaching a new course or a new grade level, the information that you receive from your tests is likely to be filled with error. Nevertheless, each year that you teach the same courses or grade level, you have a chance to improve on the measurement system that you use. Over three or four years, with a similar curriculum, the information you obtain from tests should become more reliable and more valid. In that way, your grades will be based on better information than when you started. It has often been said that a teacher who teaches the same way for 10 years does not have 10 years' experience but one year's experience 10 times. It's only when you use the information you obtain each year to design better courses in subsequent years that you actually profit from your experience as a teacher.

CONCLUSION

The better or the sweet of an instructional episode is in the assessment of student understanding. Advance organizers provide cognitive structures for new knowledge
to fit within, aiding student understanding. Relating new knowledge to existing knowledge helps too.

The schemata students bring to the instructional situation determine to a large extent what they learn. So does teacher explicitness in explaining. Teachers who are better explainers avoid vague terms, are explicit in stating what is to be learned, and model or communicate how students are to process the information. The plentiful use of examples and the rule-example-rule technique of instruction also further understanding. More effective teachers use explanatory links and verbal markers of importance when speaking. Thus good explanations occur when instructors help students to generate personal meaning, carefully communicate what to focus on, are explicit about how students should think about what they are focused on, and emphasize the most salient ideas of the content in their explanations.

Teachers use many methods to elicit information from students. Among the informal methods that teachers use is the interpretation of classroom cues. Research has found this technique to be more misleading than most teachers believe. Classroom questions, both teacher-generated and student-generated, also provide ways to assess student understanding in an informal way. A teacher's questions were shown to have instructional effects as well as utility for assessing instruction. Observation of classroom discourse leads to the conclusion that teachers should require more thought from students than they ordinarily do. Teachers can get longer and more complex answers from students if they learn to use wait time and to probe and redirect questions. Students' answers to teacher questions are the most used data from which teachers judge students' comprehension. Students' questions to teachers and classmates are another source of such information, though student-generated questions are surprisingly rare. The private teacher-student conference is a very useful way to assess understanding but may be difficult to arrange.

A little more teacher preparation and analysis are called for when semiformal assessment techniques are used. These include practice in the class and practice in the form of homework. It was noted that diagnostic questions are the best source of data for teachers to modify their instruction. Such questions can be used in both classwork and in homework. The errors students make in answering diagnostic questions provide insight into their problems of understanding. Classroom practice and homework have direct instructional effects as well. When practice in class or at home provides students with success and is directly related to the curriculum and the outcome of instruction, such practice improves student performance.

Finally, we discussed assessing student understanding by tests—a more formal assessment procedure. Teacher-made classroom tests differ in their use. Sometimes tests help in certifying competency, while at other times they aid in placement or in the diagnosis of students' problems. The kinds of information required by a teacher determine whether criterion-referenced or norm-referenced tests should be used and whether one type of test item might be better than another. Tests need to be technically sound for decisions to be made from them. Use of a behavior-content matrix in building a test ensures high content validity. Reliability should be known to judge the technical adequacy of a test and to estimate the width of the band within which a test score might be located.

Explaining and assessing for understanding are not tasks that are easily accomplished. Nevertheless, an increased reliance on research on teaching and the technology of assessment suggests that improvements in classroom instruction are possible.

FURTHER READING
Explaining for Understanding

One of the best-selling texts over the past 10 years. The first text to have as many chapters about teaching as about learning.


Questioning


Achievement Testing


REFERENCES


