RESEARCH IN LITERACY: MERGING PERSPECTIVES

Thirty-sixth Yearbook
of
The National Reading Conference

JOHN E. READENCE
Louisiana State University

R. SCOTT BALDWIN
University of Miami

With the editorial assistance of

JOHN P. KONOPAK
Louisiana State University

HELEN NEWTON
NRC Headquarters

Published by
The National Reading Conference, Inc.

1987

ASSESSMENT, ACCOUNTABILITY, AND PROFESSIONAL PREROGATIVE*

P. David Pearson and Sheila Valencia
University of Illinois

At a time when Americans are placing greater emphasis upon educational assessment and accountability, it is ironic that the nation's reading educators and teachers find themselves on the horns of a dilemma created, at least in part, by their very ability to evaluate how well children are learning to read. The tools which are intended to help the teacher in the classroom have paradoxically become the chains which frustrate individual initiative and innovation and limit professional prerogative. At the root of this problem is a notion of perceived accountability which manifests itself in many, often contradictory, ways. On the one hand, for example, there is the widespread belief among the public, local and state school boards, and many professional educators that educational accountability can be truly and accurately fixed on the basis of test results. For many teachers, on the other hand, such a belief has not contributed to their sense of professional competence and well-being; to the contrary, this belief has eroded significantly their perceptions of their prerogatives as professional educators and their ability to make or influence important decisions about educating the nation's children.

In the following essay we will pursue these points further by taking a personal and professional tour of the issues that have led us, as reading educators, to our current dilemma. We begin with a visit to three hypothetical but typical classrooms during the last three decades. Then we will examine the dangers and discrepancies which arise when the model which underlies reading research, theory, and practice comes into conflict with the model which governs our reading assessment practices, policies, and decision-making procedures. To remedy the dilemma which has arisen, we will propose an alternative way of conceptualizing the relationship between assessment and instruction. Lastly, we will close by calling for action on a research agenda in the belief that if we do not directly address the research and policy issues stemming from this dilemma, teachers' professional prerogatives will be eroded even more seriously than they have been already, with potentially grave consequences for our ability to educate our children successfully.

---

*Based on the Presidential Address and supported by the National Institute of Education under Contract No. 400-81-0030 and by the Illinois State Board of Education under Contract No. J61.

THREE SCHOOLS

The first of our hypothetical classroom visits takes place in 1964 in southern California. The classroom has 32 fifth-grade students, 15 of whom are bilingual (Spanish and English) and four of whom are monolingual Spanish speakers. The teacher has a complete set of fifth-grade basals — *Trails to Treasure,* six copies of the fourth-grade basal — *Roads to Everywhere,* a few randomly gathered old basals, a weekly supply of *My Weekly Reader,* a workbook for each of the students reading on grade level, but no ditto masters or back-up workbooks. Standardized testing consists of the Iowa Test of Basic Skills, which is given each March; the tests are scored by hand and the results are dutifully reported to the principal, who seems unconcerned about whether or not the results are reported to either students or parents.

This scenario takes place before the advent of what we have come to expect in modern basals — the ubiquitous end-of-unit and end-of-level tests. Once in a while, an occasional workbook page appears which is characterized as a progress check, but basically the teacher's manual does little to encourage the possible use of these tests to screen students for remedial activity. The teacher's decisions about what to cover in reading are guided by the school's program, as defined by the teachers' manuals in the basal program. Overall, however, the emphasis given to specific skill instruction is not, in any real sense, guided by scores on tests of any sort.

The relationship between tests and instruction in our hypothetical classroom, and during this period, was one merely of *interest:* The security of the teacher's and the principal's position and salary, and of any child's promotion was *not* on the line when they gave either standardized tests or those informal basal reader tests. And, probably most important, the reputation of the school in the public's eye was not dependent on test scores published in the local newspaper.

By the time we visit our second hypothetical classroom — a fifth-sixth grade in a suburb of Denver in 1973 — significant changes in the instruction-assessment relationship can be observed. These changes are apparent from the classroom's physical characteristics. As part of an open space school, the classroom is large but has been divided up into lots of nooks and crannies, including a whole bank of library-type carrels along one long wall. On another wall is a massive structure of cubbyholes, not unlike those used for mail distribution at colleges or public schools.

This school had become, just the year before, an *individually guided education* school and had purchased one of the then popular *skills management systems* that emerged from the mastery learning and criterion-referenced movement of the late '60s and early '70s. Every couple of months, all the students in the class take a set of criterion-referenced tests that define the scope and sequence of skills for those grade levels. Three days per week, reading class consists of students going to see their teacher to receive a list of skills that their test results suggest are weak and need to be practiced. The students then go to the big wall of cubbyholes and pick up one worksheet out of every cubbyhole identified on their skill practice list. Once students have completed the required worksheets, they are allowed to retake the tests that sent them to the cubbyhole wall in the first place. Passing one set of tests allows students to enter a new level of skill performance and begin the process again. Failure on any subtests results in remediation in the form of even more worksheets in another row of

the seemingly endless cubbyholes. There was a side-benefit, however unintentional, to this system. The other two days a week, the reading program consisted of reading stories in the basal reader and discussing them without the encumbrance of all the skills activities that typically accompany a basal lesson.

In this scenario, criterion-referenced tests of very specific skills are the foundation for individualized instruction. Instruction in a specific area is offered if, and only if, test performance warrants it. Ironically, norm-referenced, standardized tests play only a minor role in this process. Such tests are used only to evaluate whether or not the entire mastery learning, skills-management system is working generally. This system seriously erodes professional prerogative, at least in comparison to what we depicted as being representative of instructional techniques only a decade earlier. This erosion of professional prerogative is inevitable in such a system because the mastery learning mentality which undergirds it specifies the *means* of instruction very precisely, but leaves open the issue of what the ends are. Without a clear definition of these *ends,* the means become the ends of instruction. The result is that teachers and children spend their time working on activities that do little more than help the children pass the tests.

By defining reading instruction as a set of very specific components, each of which is accompanied by an equally specific test of mastery and an equally specific set of workbook or worksheet pages, there is no room or need for teacher judgment in the assessment process, in instructional decision-making, or in the delivery of instruction. Teachers become just managers, the organizers of material to be learned through repeated interactions with worksheets. Indeed, in this second hypothetical classroom, the two teachers in the team do little more than take turns monitoring the administration of tests and control of materials in the cubbyholes or making sure that students behave themselves while they are so busily engaged in so-called individualized instruction.

Our third hypothetical classroom is in a posh suburb in nothern Illinois and represents what is happening in all too many contemporary American schools. The impression one gets from talking to teachers is that of classrooms virtually inundated by tests: standardized tests, basal reader tests, teacher-made tests to give grades, and now the prospect of annually administered statewide tests of reading, language arts, and mathematics.

Due largely to a seductively attractive movement called outcomes-based education or performance-based education (see Popham, Cruse, Rankin, Sandifer, & Williams, 1985, for a clear description of this movement), standardized tests are more popular than ever before. The great selling point for standardized tests, so the outcomes-based education argument goes, is that if you agree to be accountable for broad outcomes, then you can recapture control over the day-to-day decisions about what to teach, when to teach it, and how to teach it.

Such an argument might be compelling if it were not for the continuing popularity in this district, and most other districts, of the end-of-unit and end-of-level tests in basal reader programs. In essence, standardized tests continue to determine the *ends* of instruction and the basal tests determine the *means* of instruction. Additional evidence of this control can be found in the fact that in 1987 the basal companies correlate their tests with the popular standardized measures, making for a tighter and more constraining relationship between ends and means. And, just last summer, one of the

leading test publishers undertook a project to make certain that its tests matched basal reader objectives more closely! In short, the knot becomes tighter and tighter.

The net result of this ever tighter control over both outcomes *and* processes is to leave very few decisions for teachers, principals, or central office staff, except perhaps what tests to purchase. Furthermore, unlike earlier eras, the spectre of accountability now hangs heavy in the district. It is not surprising that teachers feel compelled to worry about test performance, now that school-by-school results are published in the local newspaper. No one has been fired for having students whose performance on tests is abnormally low; however, teachers behave as if their jobs, or at least their professional reputations, are on the line. In terms of professional prerogative, the situation could not be more constraining because teachers control neither the ends nor the means of instruction. They have retained the responsibility for student performance without any authority to alter instructional programs on their own.

THE NUB OF THE PROBLEM

These three scenarios raise a number of key issues that, as researchers, teachers, and teacher educators, we need to address. We have reached a point where the threads of instruction, assessment, and decision-making are very tightly interwoven. While most of us would probably agree that this interweaving is, in principle, desirable, reality reveals that it is problematic for two reasons. First, the interplay among these components does not reflect our best understanding of the reading process. More specifically, reading assessment has not kept pace with reading theory, research, or practice. Instead of being mutually supportive, there is often disruptive tension among the instruction, assessment, and decision-making processes. Secondly, in an attempt to objectify and routinize the way data are collected and used to make decisions, the teacher has been forced out of the assessment process. So much for professional prerogative! As a field, we face a dilemma (Valencia & Pearson, 1987) with some devilish characteristics.

The first and most important characteristic of the dilemma is the conflict between our newly emerging views of reading process and reading instruction, on the one hand, and the model of the reading process that underlies our current assessment practices and procedures. Recent research (e.g., Collins, Brown, & Larkin, 1980; Pearson & Spiro, 1981) has emphasized reading as a constructive process. The reader strategically and thoughtfully uses clues from the text, background knowledge, the reading context, and other sources to gain meaning from the text at hand. At the same time, this view de-emphasizes the notion that progress toward expert reading is guided by the aggregation of component reading skills. Instead, it suggests that skilled reading, at all levels, is reflected in the reader's awareness of how, when, and why to use resources for the goal of constructing meaning. Skilled readers can use knowledge flexibly — they can both learn from the immediate reading situation as well as apply what they have learned to new situations (e.g., Campione & Brown, 1985; Spiro & Meyers, 1984).

While this model of the active, strategic reader dominates our current research base, it bears little resemblance to the model which underlies most of our reading

assessment schemes. Consider but a few of the discrepancies between what we know from research and how we assess reading:

> Prior knowledge is a major determinant of reading comprehension, yet we mask any relation between knowledge and comprehension on tests by using many short passages about unfamiliar, sometimes obscure, topics.
>
> Real stories and texts have structural and topical integrity which influence reading comprehension, yet we assess reading comprehension using short bits that rarely approximate authentic text.
>
> Inference is an essential skill for comprehending words, sentences, paragraphs, and entire texts, yet many assessments rely primarily on literal level questions.
>
> Prior knowledge and inferential thinking work together to help the reader construct meaning from the text. Because these attributes vary across individuals (and within individuals from one situation to the next) and because texts may invite many plausible interpretations, we would expect many possible inferences to fit a given text or a question. Reading comprehension, however, continues to be assessed using multiple-choice items with only one correct answer.
>
> To accomplish the goals of reading, readers must orchestrate many so-called skills, yet many of our reading assessment schemes fragment the process into discrete skills, as if each was important in its own right.
>
> Flexibility — the ability to monitor and adjust reading strategies to fit the text and the situation — is one hallmark of an expert reader, yet we seldom assess how, when, and why students alter their approaches to reading.
>
> The acid test of learning from text is the ability to restructure and apply knowledge flexibly in new situations, yet our assessment schemes rarely ask students to do so. Instead, we seem to be comfortable with tasks that seldom go beyond restating textual information.

The point is obvious but insidious: Tests continue to assume a prominent place in the assessment, instruction, and decision-making arenas, yet these very tests often represent an alternative and contradictory view of the reading process. The result is tension and confusion among professionals responsible for instructional improvement and for monitoring student progress.

This tension could easily transform itself into a kind of schizophrenia among reading program directors and reading teachers. While anxious to implement instructional practices based upon the latest research, they are plagued by the threat of low test scores. As a result, they are forced to try to integrate two diametrically opposed curricula—one based upon what is measured by the tests for which they are accountable and one based upon what they have learned from recent research. In such conflict, teachers are not likely to exercise their prerogative.

The second characteristic of the dilemma is its penchant for irony. At a time when reading tests are most in conflict with what we know about reading, they are being used more than ever. Beginning with the accountability movement of the 1970s, and moving toward the current deluge of national reports (Education Commission of the States, 1983; Fisher, Berliner, Filby, Marliave, Cahen, Dishaw, & Moore, 1978; Goodlad, 1984), reading achievement has been a major focus of most of these educational improvement efforts. And, in many cases, they have relied on students' standardized test scores as measures of effectiveness or educational quality (recall our

earlier reference to outcomes-based education). Such a reliance has led to an increased focus on testing of all sorts and at all levels.

As evidence of the increasing use of tests, we can now document at least 40 statewide competency testing programs. As one might expect, assessment efforts are not restricted to the federal or state level. One only needs to look inside schools and classrooms to find thousands of locally regulated testing programs, criterion-referenced tests accompanying basal reading programs, and the countless school- and teacher-made tests. The brief snapshot of the hypothetical school in suburban Illinois is not much of a caricature. No matter the perspective one takes on this picture, the conclusion is inescapable: The influence of testing is greater now than at any time in the history of schooling.

A third characteristic of the dilemma is its capacity to breed uneasiness in the various research communities. Our discussion in this paper is prima facie evidence of the reading research community's concern about testing. However, it is not clear that either the policy research community or the assessment research community appreciates the new-found popularity of tests. The instruction/assessment link has been vigorously debated by educators. Some sing the praises of instructional programs driven by test results (Haney, 1985; Popham & Rankin, 1981) and cite positive results to support their case (e.g., Popham, Cruse, Rankin, Sandifer, & Williams, 1985). Opponents of such testing schemes argue that tests should follow rather than lead curriculum (Berlak, 1985). They claim that overreliance on test scores leads to a narrowing of the curriculum, a tendency to teach to the test, and an emphasis on lower level, more easily tested skills (Linn, 1985). Still others (e.g., Madaus, 1985) remind us that some of our large-scale tests have become so generic and curriculum insensitive that they are virtually useless for making decisions in a school setting. While the debate continues, the inescapable truth is that assessment is a powerful force that must be reckoned with.

Dangers to the Field

The dilemmas that erode teacher prerogative are bound to create dangers. One danger stems from the impact of assessment on teachers' thinking and classroom instruction. Using results from existing measures, teachers may develop a false sense of security when they observe high scores. A close review of these tests reveals a narrow, restricted view of comprehension. Not only might teachers begin to take pride in performance that does not reflect meaningful comprehension, but they might also be encouraged to shape instruction to produce high scores on these same measures. Conscientious teachers have always wanted their students to perform well on tests; not surprisingly, they look to these tests as guides for instruction. If tests foster an inappropriate model of skilled reading, inappropriate instruction is likely to result. We end up supporting practices that promote high test scores at the expense of effective reading strategies. Furthermore, we encourage little, if any, change in instruction. Worst of all, we provide no incentive to encourage teachers to recapture their professional prerogative.

One can counter this argument by suggesting that tests, as outcome measures, were never meant to define the instructional process leading to the outcome; one can

argue further that these outcome measures only sample the vast set of possible outcomes. Such arguments, however, are specious. Educators, with good intentions, tend to teach to the test. The New York Regents exam, the California Subject A writing exam, and the European notion of the tradition of past exams all illustrate the well-entrenched propensity to teach to the test. On the positive side, look at the redirection of instructional efforts in writing, apparently induced by tests in which students actually had to write.

A second danger stems from the potential insensitivity of current assessment strategies to new instructional programs aimed at promoting strategic reading. Teachers and administrators may mistakenly interpret no, or only small, gains as indicative of an ineffective instructional program. The alternative hypothesis—that the assessment strategies are insensitive to the desired outcomes—may never be considered. Tests that are not designed to tap the strategies and thinking that are integral to skilled reading will not be sensitive to changes in these skills brought about by new instructional programs. In such a situation, educators initially motivated to take risks are not likely to continue to do so.

A third danger is that when some assessment tools become officially sanctioned, teachers tend not to rely on their own assessment skills to make important instructional decisions; ironically, of course, the data a teacher collects has the greatest potential for influencing day-to-day student learning. While most tests may be effective at indicating broad increases in reading achievement, they offer the teacher little useful information for refining specific instructional strategies. The lure of objectivity associated with commercially published tests and the corollary taint of subjectivity associated with informal assessment pushes teachers further and further away from instructional decision-making. For some reason, teachers are taught (and apparently learn) that the data from either standardized or end-of-unit basal tests are somehow more trustworthy than the data they collect each day as a part of the normal course of teaching. The price we pay for such a lesson is high since it reduces the likelihood that teachers will use data they collect themselves for decision-making within their classrooms.

If the responsibility for assessment and instructional decision-making is placed with the teacher, we will produce more capable, concerned teachers. Take this away, and we create teachers who are just managers rather than educational professionals for whom professional prerogative is synonymous with teaching. This last point is revealed most dramatically in the work of Clay (1985) and Johnston (in press). In the recently released edition of her book on informal assessment, Clay details a completely individualized and curriculum-embedded approach to assessment (i.e., there is no need for special tests or materials). In a paper in which he argues persuasively for what he calls a more naturalistic (what we will later call a more informal and locally controlled) approach to assessment, Johnston suggests a number of alternative assessment devices that teachers could use for instructional decision-making were they to abandon their reliance upon formal (norm- and criterion-referenced) tests. In both cases, the measures advocated have the virtue of being so related to instruction that assessment and instruction become indistinguishable. All of these measures also rely heavily upon individual judgment (one type of prerogative); hence, they are remedies for the kind of situations that lead to our third danger.

A FRAMEWORK FOR A COMPLETE ASSESSMENT SYSTEM

What we must do, then, is to help educators and policy-makers reconceptualize assessment. New and better tests, in and of themselves, will not solve the problem. We need to reconceptualize assessment as a framework for making decisions. Basic to this framework, we propose these assumptions:

1. Reading assessment strategies need to be based upon our best and most current models of the reading process.
2. Assessment should not drive instruction, as is currently true in American schools; instead, assessment and instruction should be so interwoven as to be indistinguishable from one another.
3. Assessment schemes which fail to capitalize upon the expertise and contextual advantage of classroom teachers ignore what may be the richest source of data for making instructional decisions.
4. There are different levels of decision-making, each with its own unique demands and, possibly, unique assessment tools. However, whatever is worthy of assessment ought to be assessable in different context levels for different purposes using different strategies.

The framework we propose for developing a complete assessment system is depicted in Table 1. The attributes of skilled reading listed in the first column represent several current working hypotheses about what it means to be a good reader—those outcomes for which students and teachers might want to be held accountable. Columns 2, 3, and 4 indicate contexts of impact at which various types of decisions are made.

It is likely that states and even large school districts will continue, at least for the foreseeable future, to collect data on large numbers of students. Hence, we will continue to need assessment strategies that use efficient and employable tools. Such tools may be useful for determining trends, but they are never likely to guide instruction at the classroom level very well unless the results are amplified by other assessment devices under the control of the school and the teacher.

To better understand what we would like to propose as an ideal for what should occur in columns 3 and 4 (especially column 4), let us consider a completely different relation between reading theory, reading instruction, and reading assessment. What would happen if we took seriously the admonition to base assessment upon a strategic model of reading? What would happen if we redefined the relationship between instruction and assessment so that it was a supportive interaction among assessment, instruction, and teacher prerogative? Readers would read to construct meaning. Every act of reading, and, therefore, every act of assessment, would be identical regardless of who was performing it. What would vary across readers, situations, and levels of sophistication is exactly how readers orchestrate available resources.

Given such a view, optimal assessment and learning would occur when teachers observe and interact with students as they read authentic texts for genuine purposes. As teachers interact with students, they would evaluate the way in which the students orchestrate resources to construct meaning, intervening to provide support or suggestions when the students appear on the verge of faltering in their attempt to build a reasonable model of the meaning of the text. This model, referred to as dynamic assessment (Campione & Brown, 1985), emanates from Vygotsky's notion of the "zone of proximal development," that region of development just far enough—but not too far—beyond the students' current level of competence such that sensitive teachers, using scaffolding tools such as modeling, hints, leading questions, and cooperative task completion, can assist learners in moving to their next level of sophistication. Instruction consists, in such a model, not of remediating deficient skills, but of using assessment strategies and observation to determine which of the potentially useful resources students have trouble using to their best advantage and then providing the scaffolding necessary to support its use. The measure of students' ability is not a score; instead, it is an index of the type and amount of support required to advance learning to a higher level of sophistication.

This scenario, in which there is no difference between reading instruction and assessment, and in which both teacher and student provide input, is an ideal. While this model is one that may never be fully integrated into large-scale tests of reading, it should be the goal of classroom and individual student assessment and instruction.

To illustrate how a single goal might be assessed across levels of decision-making, consider the domain of metacognition. Students' sensitivity to the demands of the task, audience, and situation, and their ability to vary reading strategies to meet these demands might best be assessed by observing and interacting with students while they are actually applying these strategies in real reading situations (Palincsar & Brown, 1984, 1986). We can and should attempt to measure these skills in formats amenable to

Table 1

The Relationship Between Goals, Decision-making Units, and Methods of Assessment

Hallmarks of a Good Reader	State or District	School or Classroom	Classroom or Individual
Good Readers ...			
use prior knowledge to help them construct meaning from text.			
draw inferences at the word, sentence, paragraph and text levels.			
provide many plausible responses to questions about a text.			
vary reading strategies to fit the text and the reading situation.			
synthesize information within and across texts.			
ask good questions about text.			
exhibit positive attitudes toward reading.			
integrate many skills to produce an understanding of text.			
are fluent.			
use knowledge flexibly.			

large-scale assessment. But there will always be some limitations to data gathered from group tests of metacognitive activities. These limitations are based on our observations while interviewing and testing students in a variety of situations: (a) what students say may differ from what they do, (b) strategic readers are too flexible and adaptive to allow us to capture their skill in a small sample of situations and options, and (c) for many readers, these strategies operate at an unconscious, automatic level inaccessible to verbalization or even reflection. In short, here is a case in which large-scale assessment may prove moderately useful for some very limited purposes and decisions; however, the assessment strategies that really count are likely to occur at the classroom or individual level.

What we are really saying is that our 80-year history of assessment in America has focussed only upon column 2 in Table 1 (large-scale assessment). The only serious attempt to deal with decision-making needs in columns 3 and 4 was the mastery-based criterion-referenced systems that arose during the 1970s. The problem is that these systems applied the principles and assessment techniques associated with large-scale assessment and decision-making to a situation that demanded fewer constraints and much greater flexibility. The problem with both the norm- and the criterion-referenced versions of commercially available tests is that they have assumed that teacher judgment is unnecessary in making decisions. To achieve our goal of a balanced system of assessment, educators must commit themselves to at least two tasks: (a) they must assume that any data from commercial tests is information that they have to interpret in concert with other information they possess about schools and students, and (b) they must begin to collect data in their own situations so that they can build portfolios of information about students, classrooms, or schools.

One could argue that an opportunity to achieve this goal has always existed and that teachers could, if they desired, avail themselves of a wide range of assessment strategies, including strategies that emphasize rich data bases gathered in interview-like situations. However, this argument fades rather quickly in light of the low status accorded to informal assessment and teacher judgment.

A CALL FOR ACTION

As surely as we see the need for more natural assessment opportunities, we must not forget that standardized, norm-referenced tests are still the most prevalent type of testing in our schools. They deserve our immediate attention. We need research to develop and evaluate new assessment techniques that are both consistent with our understanding of reading and reading instruction *and* amenable to large-scale testing. Unless we can influence the shape of large-scale assessment, we may not be able to refocus assessment at all, thereby losing our opportunity to expand the framework of assessment to include more naturalistic and instructionally valid approaches.

New Research Efforts

The development of new techniques is only one part of this process. And that has been the focus of our work in Illinois (see Valencia & Pearson, 1986) and four

colleagues in Michigan (Wixson, Peters, Weber, & Roeber, 1987). There is also a great deal of research still to be done on these issues. If we are ever to develop anything like the assessment system we have argued for in this paper, critical research efforts need to be made in several areas.

Validity efforts. We need different sorts of validity indices for paper and pencil measures. First and foremost, we should recognize that we will develop the best assessment devices, be they formal or informal, as a result of attempts to establish the validity of the theoretical constructs that underlie the tests we create. Such efforts must constantly bear in mind the old truism that reading is a complex process. It is unlikely that a simple test of reading will ever stand the scrutiny of construct validation.

Second, we should outlaw the practices of concurrent validation that permit Test 1 (usually an experimental form) to be validated by suggesting that it correlates highly with Test 2 (usually a widely used standardized test). Such practices only perpetuate the conventional wisdom responsible for where we find ourselves today. If we permit any form of concurrent validation, the criterion against which a candidate test is compared should be a measure of reading which has a high degree of ecological and construct validity.

Third, we need to take the whole issue of instructional validity seriously. We certainly do not want to permit the legal definition of instructional validity, as in the Debra P. case. The issue in that case was whether or not the test tested what was taught in the schools (much like what we have traditionally referred to as face or curricular validity). We need to define instructional validity in terms of instructional sensitivity; that is, a test will be regarded as instructionally valid to the degree that it is sensitive to the growth that we know will occur when we engage students in certain instructional activities. In essence, what we need to do is to turn our usual experimental approach on its ear. Usually we assume the validity of some outcome measure and then evaluate the efficacy of competing instructional approaches in terms of how well students perform on that common measure. In the approach we are suggesting, we would assume the validity of the instruction and evaluate the validity of competing measures of the process being learned.

System exploration. If we are ever to infuse informal (what Johnston wants to call more naturalistic) measures with the degree of credibility necessary to encourage teachers to use them, we need to conduct some very convincing research demonstrating the usefulness of approaches to assessment that are so informal as to be indistinguishable from instruction.

Literacy experiences. For nearly 70 years, we have assumed that the ability to perform some cognitive task was the ultimate in reading assessment. And given the individual differences milieu in which reading tests arose at the time of the First World War, it is not surprising that a cognitive bias prevailed. However, other, less direct and less obtrusive measures, particularly related to programs rather than to individuals, have been overlooked. For example, if you ask yourself, how do you know when your reading program is working?, you might come up with indices like these:

When subscriptions to the local newspaper rise.
When the checkout rates at the local library rise.

When book sales at the local book stores, particularly children's book sales, rise.
When surveys indicate that the amount of time children read voluntarily at home (or in school for that matter), increases.
When attitudes toward reading become more positive.
When students indicate they understand what their teachers are trying to do.
When businesses stop complaining that they have to teach literacy skills to our graduates.
When teachers demand and regain their professional prerogative.

It is only an historical accident that reading tests became established during a period in which educators sought to measure individual differences scientifically and objectively; in another era we might have developed a very different set of measures of reading effectiveness.

Understanding the present. Much of what we have claimed about the current uses of tests is based upon what even the most ardent ethnographers would refer to as informal evidence. We need to understand more fully both the dynamics and the consequences of the public's demand for accountability and assessment. To this end we need good ethnographies of the real and perceived uses of assessment devices conducted at the classroom, school, and district levels. Secondly, we need some intervention work on the assessment-instruction link. For the past several years we have too easily resigned ourselves to the supposed truth that assessment does indeed drive instruction. If such an assertion is true, an assessment system built on a different model of reading should result in categorically different kinds of reading instruction. If this hypothesis is untrue, we need to see if things can be turned around.

Reporting information. We would probably all concede that, as researchers, we are sometimes poor communicators. But if we think our communication about research is inadequate, our communication about assessment (what scores really mean) is abysmal. We need research, most likely done by experts in communication and information dissemination, that will ultimately improve the way in which we report assessment data to various audiences. One of the biggest problems we will have to address is how to get the general public to understand the limitations of test scores (and assessment data generally); at present, the American public is unfortunately all too credulous about standardized test scores.

Creative efforts. Above all, we need for researchers and educators at all levels to think creatively about formats and systems for assessment at all levels of decision-making, be it at the level of the state, district, school, classroom, or individual. This is a time to give rein to our imaginations and our problem-solving capacities.

Development Efforts

While researchers begin to explore the theoretical and psychometric aspects of new large group formats and techniques, we must also all work to develop the disposition and specific assessment techniques needed to answer the varied questions posed by people charged with decisions at different levels. Our goal, as reading educators, should be to develop valid, reliable, and usable strategies to be used at all levels of decision-making. Only if we are able to fill *all* the cells of the chart in Table 1 with conceptually sound assessment strategies will we approach our goal of equipping

educators with a portfolio of assessment strategies that they can use to make and then implement, sound decisions.

Finally, there is the issue of what stance we shall take toward improving or reforming the current system of assessment. There are those among us who will say that the current system is so corrupt and compromised that it doesn't deserve any attempt to reform it. Toss it out, they will say, and let's start over from scratch. Then there are those among us who will choose, at least for the moment, to try to reform the system from within. Both positions, it seems to us, have their merits at this point in the history of assessment, and either course of action is preferable to clinging to a moribund status quo.

We close with a telling quotation from a student of medicine. In 1979, Stanley Joel Reiser observed:

> If physicians in general come to accept a fundamentally mechanical view of human beings, in a world that is more and more enamored of technology, the prospect for the future is extremely disquieting.... Machines inexorably direct the attention of both doctor and patient to the measurable aspects of illness, but away from the "human factors" that are at least equally important.... Technologies that improve accuracy, and centralized organizations that enhance efficiency and improve security, are essential factors in modern medicine. Yet accuracy, efficiency and security are purchased at a high price when that price is impersonal medical care and undermining the physician's belief in his own medical powers. To be free to develop his medical skills to their highest point, to increase what is, despite these problems, a positive balance of benefits over harms, today's physician must rebel. He can use his strongest weapon—a refusal to accept bondage to any one technique, no matter how useful it may be in a particular instance. He must regard them all with detachment as mere tools, to be chosen as necessary for a particular task. He must accept the patient as human being and regain and reassert his faith in his own medical judgment. (pp. 229–231)

The analogy between medicine and teaching, while potentially misleading in some cases, is appropriate here. What Reiser has to say about doctors reserving a certain degree of professional prerogative is equally true of teachers. Both are professionals who must, by definition and of necessity, make decisions on the basis of incomplete or ambiguous evidence. Reiser's observation that doctors cannot hide behind the cloak of technology to explain their errors in judgment is a good message for teachers and testing technocrats to remember. With these thoughts in mind, let us redirect our conceptual and technological efforts regarding assessment to activities that acknowledge rather than ignore teachers' professional prerogative. To do less is to insure a system of assessment and instruction that, while tolerated by most, will be admired by few, least of all ourselves.

REFERENCES

Berlak, H. (1985). Testing in a democracy. *Educational Leadership, 43,* 16–17.

Campione, J. C., & Brown, A. L. (1985). *Dynamic assessment: One approach and some initial data* (Tech. Rep. No. 361). Urbana: University of Illinois, Center for the Study of Reading.

Clay, M. M. (1985). *The early detection of reading difficulties: A diagnostic survey with recovery procedures* (3rd ed.). Exeter, NH: Heinemann.

Collins, A., Brown, J. S., & Larkin, K. M. (1980). Inference in text understanding. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 385–407). Hillsdale, NJ: Erlbaum.

Education Commission of the States (1983). *Calls for educational reform: A summary of major reports.* Denver, CO: ECS.

Fisher, W., Berliner, D., Filby, N., Marliave, R., Cahen, L., Dishaw, M., & Moore, J. (1978). *Teaching and learning in elementary schools: A summary of the beginning teacher evaluation study.* San Francisco, CA: Far West Regional Laboratory for Educational Research and Development.

Goodlad, J. I. (1984). *A place called school: Prospects for the future.* New York: McGraw-Hill.

Haney, W. (1985). Making testing more educational. *Educational Leadership, 43,* 4–13.

Johnston, P. (in press). Steps toward a more naturalistic approach to the assessment of the reading process. In J. Algina (Ed.), *Advances in content based educational assessment.* Norwood, NJ: Ablex.

Linn, R. (1985). Standards and expectations: The role of testing (summary). *Proceedings of a National Forum on Educational Reform* (pp. 88–95). New York: The College Board.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66,* 611–617.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 2,* 117–175.

Palincsar, A. S., & Brown, A. L. (1986). Interactive teaching to promote independent learning from text. *The Reading Teacher, 39,* 771–777.

Pearson, P. D., & Spiro, R. J., (1981). Toward a theory of reading comprehension. *Topics in Language Disorders, 1,* 71–88.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement driven instruction: It's on the road. *Phi Delta Kappan, 55,* 628–634.

Popham, W. J., & Rankin, S. C. (1981). Minimum competency tests spur instructional improvement. *Phi Delta Kappan, 62,* 637–639.

Reiser, S. J. (1979). *Medicine and the reign of technology.* New York: Cambridge University Press.

Spiro, R. J., & Meyers, A. (1984). Individual differences and underlying cognitive processes. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 471–504). New York: Longman.

Valencia, S. W., & Pearson, P. D. (1986). *Reading assessment initiatives in the state of Illinois.* Springfield, IL: Illinois State Board of Education.

Valencia, S. W., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher, 40,* 726–732.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). New directions in statewide reading assessment. *The Reading Teacher, 40,* 749–755.