Exemplary testing programs emphasize improvement of teaching and learning rather than bureaucracy. Testing is on the increase! Business for commercial test publishers is up and new companies are eyeing the market, ready to jump into a business that looks like it is firmly standing on the up escalator. Nearly every large education reform effort of the past few years has either mandated a new form of testing or expanded uses of existing testing.

Thus Chris Pipho (1985) of the Education Commission of the States began a recent article on the education reform movement. Standardized testing in schools clearly seems to be on the increase. Exactly how much no one knows for sure—partly because so many different agencies administer tests. Many school districts have their own testing programs. Most states require schools to give special state-sponsored tests. As part of the strings attached to federal money, the U.S. Department of Education often requires the testing of children enrolled in federally funded programs. And if a kid makes it through high school, about the only way he or she can get into a four-year college without taking a college admissions exam would be by going over the ivy-covered walls after dark.

Why all the testing? Much of it is described as “educational testing,” but most of it has far less to do with teaching and learning than with school administration and educational bureaucracy. Testing has long been associated with reform—not just in the current wave of enthusiasm for educational reform, but also in reforms of the 20s, 30s, 50s, and 60s. In some of the earlier episodes testing has been not just an instrument, but also an object, of reform (Haney, 1984, recounts some of this history.) Yet considerable experience, both in the U.S. and abroad, has repeatedly shown that when emphasis is placed on test results, the beneficiary may be the test scores rather than more general learning.

After briefly describing why the phrase “educational testing” may be mainly a malapropism, I recount an effort undertaken at the suggestion of Educational Leadership to locate and describe promising testing programs in U.S. schools, four are profiled here. Finally, I suggest some general conditions that seem to contribute to making testing more conducive to teaching and learning.

“Educational Testing” as Malapropism
“Educational testing” is a pretty common term. Most of the standardized testing going on in schools is, after all, conducted in the name of education. There is, nevertheless, much that could be criticized about the limits of these tests. Indeed, in recent years, muckraking journalists and some educational reformers have regularly attacked both testing and the testing company that uses the phrase “educational testing” in its name. In 1980, for instance, one of Ralph Nader’s organi-
lications released The Reign of ETS: The Corporation that Makes Up Minds (Nairn, 1980). In 1981, Andrew Strenio's The Testing Trap was published. In the same year, Goodbye to Excellence: A Critical Look at Minimum Competency Testing by Mitch Lazarus appeared. And this year, David Owen's None of the Above: Behind the Myth of the SAT was released with considerable publicity.

One need not be a dyed-in-the-wool social Darwinist to recognize that standardized testing is increasing because it serves some important social functions. Standardized college admissions tests clearly do help selective institutions throughout the country. (In a lot of places, though, it is clear that admissions tests are only ritually required, to keep up appearances rather than to inform admissions decisions. Some people—see Crouse, 1985, for example—have sharply questioned how much college admissions tests contribute to improving selection decisions.) Standardized tests also help federal and state government agencies compare the outcomes of various government-funded educational programs (such as Chapter I programs) administered in different schools. They clearly can help school systems decide how to place students in different educational programs (for example, whether to place students in special education programs or, at the high school level, whether students should be placed in academic, regular, or vocational tracks). The appropriateness of such uses is, of course, open to debate, but an easily overlooked point is that these uses have more to do with the bureaucratic nature of schools than the real stuff of education, namely individual teachers teaching and individual students learning.

To realize how irrelevant most testing programs are to promoting individual students' learning, all one need do is compare them with almost any learning theory. One nearly universal desideratum in all learning theories is that to learn, an individual needs to receive rapid and specific feedback on what is attempted. However, this simply does not happen with most school testing programs. Typically, students receive only an aggregate score (percent correct or percentile score) after several weeks' delay (while answer sheets are sent away to be scored). From the point of view of promoting learning, there are other deficits to most standardized testing, but this is one of the most clear-cut: students do not receive rapid and specific feedback on their work.

Second, most standardized tests have a very uncertain relationship to the specific teaching and learning that occurs in particular schools and classes. In this regard, I recall a conversation nearly ten years ago with a young woman in a South Carolina schoolyard. Her class had recently completed one of the big nationally normed achievement batteries. When I asked her how she had liked it, she said it was okay, but then added plaintively, "It had division problems, and we haven't done division yet!" As Papam (1978) put it several years ago in making the case for criterion-referenced tests, "Because of the way they are constructed, norm-referenced tests can result in unrecognized mismatches between what is being taught and what is being tested." The prospects for finding a criterion-referenced test matched to what is taught in a particular class or school are perhaps better than those with regard to norm-referenced tests, but it seems almost inevitable that any externally developed test is going to have a fairly imperfect fit with the educational goals of a particular teacher or school. This is not to say that educational tests should cover only what has been taught—but the difference surely should be known when we are testing things that have or have not been taught. Moreover, the aims of education extend far beyond things that can be tested—however the tests are referenced.

A third factor indicating the limited educational utility of most standardized testing programs is the frequent concern for keeping them secure. When testing programs have large consequences for kids, teachers, or schools, considerable pressure can build up to boost test scores regardless of the method. Scandals over cheating on tests hit the front pages with remarkable regularity. New York and San Diego are two places, for instance, where cheating on tests has made headlines in recent years. So with the increase in testing as part of recent educational reform efforts, there seems to have been a concomi-
tant increase in concern for test security. Pipho (1980, p. 19) reports, for example, that "a number of aspects of the reform movement in South Carolina are built on the increased use of test scores, and this has resulted in some breaches of security." As a result, the South Carolina legislature is considering passage of a bill that would provide for fines of up to $1,000 or as much as 90 days in jail for teachers, principals, or others convicted of permitting students access to test questions, releasing test booklets or answer keys, or otherwise breaching test security.

Ninety days in jail for breaching test security! How does this square with the theory that good teachers should make clear to students their expectations for learning? One reason for the seeming discrepancy is that virtually all testing programs employ the multiple-choice format, and scoring well can be accomplished simply by memorizing the answers even if you can't actually solve the problems for which the multiple choices are offered.

One could continue on this topic—the limited educational utility of most standardized testing programs—but at some point educators have to turn to the topic of how testing programs can be made more educationally useful.

**Survey of Educationally Noteworthy Testing Programs**

Editor Ron Brandt of *Educational Leadership* asked me to write an article on noteworthy or promising educational testing practices in school systems around the country. After all, he suggested, criticism will only get us so far. What would be more useful would be an account of unusual testing programs that provide examples of educationally useful testing. After listening to his persuasive argument, I hesitantly agreed to tackle the project. My hesitation stemmed not from any doubts about the potential importance of the issue—but from the immensity of the task. About five years ago the National Institute of Education funded a quarter-million dollar project to locate and describe exemplary evaluation and testing practices in local school systems around the country. The project had several full-time staffers and lasted over a year.

There I was agreeing to do much the same thing in my spare time and with no budget. So in December 1984, *Educational Leadership* published a notice requesting nominations of noteworthy educational testing programs in schools and districts around the country. Respondents were asked to provide a brief rationale for their nomination and the name of a contact person. Thirty programs were nominated, and ten inches of documentation were forwarded for perusal. Among the reasons for nominations were that tests were being used for "curriculum evaluation"; test results were being used to track student performance "longitudinally, over time"; a testing program was being used to identify "possible strengths and weaknesses"; "local writing norms" were being developed; and schools were

---

**Item Response Scaling, Computerized Adaptive Testing, and the Portland Achievement Level Testing**

In traditional test analysis, items are characterized statistically mainly in terms of two indices: difficulty (such as the percentage of a sample getting the item correct) and discrimination (a measure of the extent to which results for the item tend to agree with results for other indicators or the same characteristic being measured, most commonly how an item behaves when administered with a specified collection of other items given to a specified sample of test-takers.

Item response scaling is becoming increasingly popular in testing circles because it promises to make both test construction and test administration more flexible by characterizing items largely independently of other items and particular samples of people who take the item—provided, of course, that some key assumptions can be met. Item response theory uses logistic statistical functions to characterize items (using item characteristic curves) in terms of one, two, or three item "parameters." The one-parameter model is often called the Rasch model after the Danish mathematician who pioneered in the development of the model.

Many of the ideas and techniques used in item response scaling are referred to as latent trait theory or latent trait scaling, but because of the potential dangers of making inferences on the basis of test scores about individuals' "latent traits," I prefer the more pragmatic phrase "item response scaling."

Item response scaling has found a natural application in computerized adaptive testing. In this form of testing, students are given test items via computer and, depending on their performance on the first dozen or so items, subsequent items presented are either harder or easier. In theory, this sort of procedure may allow more efficient testing in that individuals' performance levels on well-specified traits may be estimated with a given level of accuracy by administering a smaller number of items.

For more information on how the Portland Public Schools are using item response scaling and experimenting with computerized adaptive testing, write to Walter Hathaway, Evaluation Department, Portland Public Schools, P.O. Box 3107, Portland, Oregon 97208-3107.
machine-scoring their own tests.

I selected four cases for documentation in this article because the information accompanying their nominations seemed especially intriguing and because they represent a range of diversity with respect to both geography and testing practices.

The four cases are:

- Portland, Oregon, Public Schools and their locally developed Rasch-scaled testing program and experimentation in computerized testing
- Orange County, Florida, Public Schools and their use of the Degrees of Reading Power Test in reforming reading instruction at the secondary level
- Pittsburgh Public Schools and their "Monitoring Achievement in Pittsburgh" project
- The Prospect School in North Bennington, Vermont, which quite purposely uses no standardized testing.

My inclusion of the Prospect School calls for a bit of explanation. In her response to the Educational Leadership call for nominations, Claryce Evans of the Boston Public Schools wrote:

I suggest you add a category, noteworthy schools with no testing programs, and I nominate the Prospect School of North Bennington, Vermont.

Her nomination proved irresistible. For one thing, she suggested implicitly that my inquiry was somewhat wrongheaded, leaving out an important category. Second, it was fairly unusual in that a person working in one school system nominated a school outside that system. So I inquired further and found that the Prospect School clearly makes for an interesting contrast when it comes to testing.

I wrote each of the four accounts after reading medium- to high-sized piles of papers on the case and after telephone interviews with one or two people in each location. I was unable to visit any of the schools to get a ground-level view of the testing program in action. (The nonexistent "budget" for the project wouldn't quite cover travel costs.)

"To realize how irrelevant most testing programs are to promoting individual students' learning, all one need do is compare them with almost any learning theory."

The Portland, Oregon, Public School System

The Portland Public School (PPS) System is a large operation, serving a student population of 51,000 in 89 schools with a staff of 2,500 certified teachers and a total annual budget of $237 million. Testing in Portland is somewhat unusual in that the district mainly uses the PPS Achievement Level Test, which, though it covers the subject areas of reading, mathematics, and language usage—the same areas as most commercially produced achievement batteries—is homegrown. The purposes of the program are:
- Evaluate student progress
- Provide diagnostic information for instructional purposes
- Inform student grouping and placement decisions
- Identify goals for emphasis
- Evaluate instructional programs
- Provide information for planning
- Provide accountability to the school board and community.

To serve this range of purposes, the PPS system has developed what Walter Hathaway, the PPS Director of Research and Evaluation, calls a "layered approach to reporting," so that data from the same test can be reported in a variety of ways to serve these diverse purposes. Apart from having a fairly sophisticated approach to analyzing and reporting test data, the PPS testing program would, so far, not seem highly unusual. However, three things about it are unusual: (1) the way the district has used item response theory to construct its achievement tests; (2) the priority it gives to instructional use of test results as opposed to administrative ones; and (3) the experimentation that the system shows in refining its testing program.

On the first point, Portland was the country's first school district to use item response theory to develop an achievement test keyed to its own curriculum objectives. Developments along these lines began in the 1970s, with elaboration of systemwide goals across grade levels and experimentation with Rasch scaling. (The Rasch model is a particular form of item response or item characteristic curve scaling; see box, p. 6, for a brief description.) The first PPS tests using Rasch scaling, in reading and mathematics, were introduced in 1979. Language usage was added in 1999. Item characteristic curve scaling also brings several practical benefits. Describing the characteristics of test items independently of the particular collection of items in which they are embedded when administered allows much greater flexibility in grouping together different items for administration, while still allowing reporting of results on a common scale. This makes it much easier to organize "functional-level" testing—that is, testing that is geared appropriately to students' levels of performance. Second, it is a much more economical way of relating performance on different levels of a test to a common scale, thus allowing the charting of progress over time. (In contrast, "vertical equating" for tradi-
The Degrees of Reading Power Test and How It Is Used in Orange County

The Degrees of Reading Power (DRP) test is designed to measure strictly reading comprehension. Traditional reading comprehension tests typically provide a reading passage and then present a series of questions based on the passage. However, so-called passage dependency studies have shown that these kinds of “reading comprehension” questions often can be answered correctly more frequently than predicted by random guessing even when the passage on which the questions are ostensibly based is not provided.

The DRP test seeks to get around this problem by using a modified cloze procedure. It uses reading passages from which certain words have been deleted. The test-taker is expected to read the passage and supply the missing word from among five choices for each deletion. Each word choice would make sense in the sentence for which it is offered if the sentence were read in isolation. The correct choice is apparent only from the context in which it appears in the reading passage.

The test is scored on a scale of DRP units, ranging from 15 to 100. The same scale is used by the College Board to rate the difficulty of “readability” of textual material, using a formula derived from the Bormuth Mean Cloze Readability formula. It takes into account characteristics of text such as word length, sentence length and complexity, and proportion of commonly used words. The DRP Readability Analysis Service offers DRP ratings for hundreds of textbooks and other instructional materials. This means that students’ DRP scores can be directly compared with the difficulty of textual materials.

The DRP program suggests that students’ reading comprehension can be interpreted in three different ways:

- **Independent** is the level of text difficulty at which a student is able to read fairly easily, perhaps for personal enjoyment or independent learning.
- **Instructional** is the level of difficulty at which a reader may need some assistance.
- **Frustration** is the level of difficulty likely to confuse a reader because it is too difficult to allow comprehension.

Further information on the DRP program may be obtained from DRP Services, The College Board, 888 Seventh Avenue, New York, New York 10010.

When contacting the College Board about the DRP test, be sure to ask about the technical manual.

For information on how the DRP is being used in Orange County, write to Joy Monahan, Orange County Public Schools, P.O. Box 271, 434 N. Tampa Avenue, Orlando, Florida 32802.

The Orange County, Florida, Public School System

A second intriguing testing program belongs to the Orange County, Florida, Public School System, and in particular its use of the Degrees of Reading Power (DRP) test to help reform reading instruction at the secondary level.
level. Somewhat larger than the Portland school system, Orange County serves 80,000 regular students in 150 schools with an annual budget of $328 million. The system uses a variety of tests, including the Comprehensive Test of Basic Skills and Florida's State Student Assessment Test (parts I and II), but what attracted my interest was how they seem to be using the Degrees of Reading Power (DRP) test to help focus attention on and improve secondary reading instruction.

The DRP is a relatively new test marketed by the College Board. Developed under a grant from the Carnegie Foundation by Bert Koslin and associates working with the New York State Department of Education, the DRP differs from most standardized reading tests in two respects. First, the DRP focuses exclusively on reading comprehension. Unlike many other reading tests, it does not treat vocabulary and word knowledge as separable subskills of reading. Second, the DRP uses a scoring system that allows student performance to be compared with the difficulty of textual materials. The DRP scale is apparently much more useful in this regard than crude "grade-level" estimates of readability. (See box, p. 8, for more information on the DRP.)

The DRP was introduced into Orange County in 1983 as part of an effort to improve reading skills at the secondary level. Secondary reading specialists work with academic subject teachers to evaluate both the reading demands placed on students and their reading comprehension skills. Results of tests administered in several classes in the first year revealed that texts were simply too difficult for students to comprehend. Instead of finding easier texts or pulling out weak reading students for extra instruction, the specialists and teachers worked to develop students' reading skills in the context of academic subjects. According to Joy Monahan, the Orange County program consultant in secondary reading, the activities used included: more focus on students' background knowledge, previewing and predicting, self-monitoring and interaction, guided and independent practice, and feedback and review. In other words, Orange County is using a variety of methods to make kids more active, self-critical readers.

What most attracted my attention to the Orange County reading initiative, though, was the apparent way in which information from the reading test is being used as a spur to help make subject-matter teachers more actively involved in evaluating and improving their own teaching. According to Monahan, the use of the DRP to help spot discontinuities between reading skills and instructional materials has led to a sort of "action research" that spawns teachers' sense of ownership over strategies for teaching reading. In other words, the DRP seems to be serving as a spur not only to make students more active readers in academic classes, but also to make teachers more active and self-critical about their own work. These are obviously the aspects of which Monahan is most proud—but she adds that reading scores have gone up too, and that's always "an attention getter in test-happy Florida."

**The Pittsburgh, Pennsylvania, Public School System**

Pittsburgh is another large school system, serving 41,000 students in 81 schools, with an instructional staff of 2,700 and a $29.5 million annual budget. In 1980, Pittsburgh hired a new superintendent, Richard Wallace. A wide-ranging needs assessment followed, out of which the need to increase student achievement emerged as a high priority. The Monitoring Achievement in Pittsburgh (MAP) project evolved as a primary vehicle for pursuing this goal. The Monitoring Achievement in Pittsburgh (MAP) project evolved as a primary vehicle for pursuing this goal.

MAP is officially described as "an instructional testing system designed to increase student achievement in basic skills." One point that attracted my initial interest is that "basic skills" include not just the three R's but also critical thinking. The MAP project involves periodic testing throughout the school year in mathematics, grammar, composition, reading, and critical thinking (next year science will be added). Tests were developed follow-

"From the point of view of promoting learning, there are other deficits to most standardized testing, but this is one of the most clear-cut: students do not receive rapid and specific feedback on their work."
"... it seems almost inevitable that any externally developed test is going to have a fairly imperfect fit with the educational goals of a particular teacher or school."

Monitoring Achievement in Pittsburgh

Relatively little on Monitoring Achievement in Pittsburgh (MAP) has been published outside of Pittsburgh. However, LeMahieu (1984) presents a brief account of MAP and results of an interesting analysis showing that improvement in student test scores in Pittsburgh on the California Achievement Test (CAT) have been larger on portions of the CAT that overlap with the content of MAP. LeMahieu also offers some important cautions about the possible long-term effects of a program such as MAP, namely that:

1. "Teachers may come to accept the objectives of the monitoring program as the sole content of instruction in the domain."
2. Over time there may be a "loss of residual learning outside the MAP content."
3. Increasing monitoring testing in additional areas competes with "an extremely important and limited resource: instructional time."

For more information on MAP, write to Paul LeMahieu, Board of Public Education Administration Building, Bellefield at Forbes Avenue, Pittsburgh, Pennsylvania 15213.

Reference


and debate about what are the key systemwide goals and idea sharing about how to teach the most important things.

Second, if some three to six times a year you are testing—or monitoring—mathematics, grammar, composition, reading, critical thinking, and science, aren’t you over-testing kids? LeMahieu’s answer went roughly as follows. MAP doesn’t attempt to rigorously test skills in each area. The monitoring tests have only one item per objective, so obviously you can’t make any inferences about individual students based on one item. The intent is simply to provide a general check on their progress in priority areas. Moreover, LeMahieu explained, each MAP test takes only 18 to 20 minutes. As MAP tests are introduced in new areas, the frequency of testing in other areas is cut back so that total time devoted to testing is not increased.

The MAP testing is not all of the multiple-choice item format. In the composition and critical thinking areas, students respond with short essays. LeMahieu says that he wouldn’t necessarily want to hold up the assessments in these areas as psychometrically ideal, but they do seem to have promoted more attention, concern with, and time devoted to writing. This was, he suggested, something that didn’t seem to be happening in any coordinated way before the MAP writing instruments were introduced.

What most impressed me about MAP was the extent to which it seems to be viewed primarily as an instructional, rather than administrative, tool. Two indicators were the way in which teachers were involved in its development and the way in which some items were included because of their perceived educational importance, despite doubts about the extent to which they could be measured according to prevailing standards of psychometric quality. One specific comment by LeMahieu convinced me of this point. He recounted that one of the first things he did after becoming Director of Testing and Evaluation (LeMahieu came to this position in 1982, after MAP had been begun under Wallace).
was shred reports comparing schools on the basis of MAP results. Why? Because once you make an instrument into a tool for evaluating teachers and schools, you automatically communicate that it is not their instrument; a source of useful but fallible information to inform their professional judgment.

What have been the results of using MAP as part of the broader school improvement program in Pittsburgh? LeMahieu answered that, among other things, test scores have been going up. But test scores always go up in these kinds of programs, so I don't put much stock in that. Perhaps more important is that high school teachers and teachers at other transition points are reporting that kids are coming in better prepared in the priority areas Pittsburgh has focused on. Also, parents seem to be getting more involved. Though the lack of parent involvement tends to be a tremendous problem in large city school systems, Pittsburgh educators have been impressed with the extent to which MAP seems to help communicate the schools' expectations to parents. It has been somewhat surprising, LeMahieu said, how it pleases parents simply to be told, in concrete ways, what the schools expect of students.

So, my final question to LeMahieu was, "You have developed an elaborate process for identifying system-wide educational priorities and some checks to help inform teachers and parents about how well kids are doing in these priority areas of learning. Isn't this something that should be going on anyway?" "Well," LeMahieu responded, "ideally there should be common expectations about what should be learned and, as a normal part of teaching, teachers should be checking on how kids are learning. But in a large system with many kids moving from one school to another every year, it is easy to lose sight of common expectations and checks—not just tests—to help judge how well expectations are being met. Although it's unclear how large a system has to get before those common expectations tend to break down," he continued, "it definitely seems to be less than 40,000 kids and 80 schools."

The Prospect School of North Bennington, Vermont

My last case is the outlier—the school nominated for this survey because it has no standardized testing. Prospect School is a small private school in independent-minded Vermont, part of a larger nonprofit organization. The Prospect Archive and Center for Education and Research, headed by Pat Carini. The school has but three teachers and 66 students, aged from roughly 4-1/2 to 13. With only three teachers, it is not graded; instead children are grouped into West, East, and South groups, based loosely on age.

Prospect School was founded in 1965 on the "generative principle" that the school's practice should be "guided by regular and frequent observation of students; as they engage in the material, instructional, and social content of the classroom," it is based on the belief that purpose, value and knowledge are occasioned when individuals give shape and order to their own experience, to the stuff of the world and of culture, and when they form connections between their own experience and that of others, and the common inheritance. Actions, events, and information take meaning from context and relatedness; meaning, importance and worth are matters of relationship and engagement which do not arise in isolation or when experience is fragmented.

According to Carini, the school used standardized tests for a few years but has since purposely avoided them. Why? "The school does not use grades, standardized tests or other external standards because we are committed to more comprehensive ways of evaluating children's progress." The "more comprehensive" means of evaluation include close observation of students and classes, written descriptions of children and the life of the school, and documentation of children's work, including collecting samples of their writing, drawings, and other projects. Records are kept in narrative form (not, for example, by using checklists), and may contain accounts of what children like to do, how they work in groups and with other children, and their "involvement with formal subject matter." Reading through samples of Prospect descriptions of children, including Observation Reports, Parent Reports, and Transition Reports (used when children leave Prospect to go on to other schools), one does gain a feeling for the child as an individual—his or her likes, dislikes, and personality, for example—that never comes through in test reports or typical school records.

Prospect School staff members meet regularly—at least once a week—to review records or issues. These reviews, also called reflective conversations, may focus on a particular child, a curriculum issue, or a more general issue or problem. These events are intended as collegial reviews in which staff members, and sometimes outsiders, attempt to deepen their understanding of particular children, pedagogical practices, ethical issues, or sometimes just school procedures.

But, I asked Carini, in this sort of a process-oriented approach to evaluation, couldn't standardized test results be useful, providing another perspective on children and their learning? The notion that tests can be used in some positive way rather misses the point, she responded. Standardized testing inevitably engages in abstract comparisons and in so doing diverts...
Documentation and Description at the Prospect School

The Prospect School of North Bennington, Vermont, is part of the larger nonprofit Prospect Archive and Center for Education and Research. Established in 1965, the Prospect School is an alternative school for children aged 4 1/2 to 13, with instruction "individualized in the tradition of open education." For more information on the school and examples of its documentary procedures, write to Dirck Roosevelt, The Prospect School, North Bennington, Vermont 05257.

Information on the educational philosophy undergirding the Prospect School, as well as its methods of observation, description, and documentation, have been published in the monograph series of the North Dakota Study Group on Evaluation. For information on this series, write: North Dakota Study Group on Evaluation, Center for Teaching and Learning, University of North Dakota, Grand Forks, North Dakota 58202.

The Promise of Educational Testing

What can be concluded from this brief review of noteworthy testing programs—or in the last case, the "noteworthy school with no testing program"? One minor point is to beware the silver-tongued editor who persuades you to try to do the impossible in 20 pages. Many of the other testing programs nominated surely would have been worth closer scrutiny. One cannot reasonably, in a spare-time effort, hope to thoroughly survey potentially noteworthy testing programs. Moreover, in such a limited space, one cannot do justice to even one of the four initiatives described. For this reason, I have tried to provide references for the interested reader on sources of further information.

More generally, one important point emerges from this review. The term "educational testing" often is used interchangeably with "educational measurement." This usage tends to focus attention on the technology of testing: the scoring, scaling, and reporting. However, what is going on in Portland, Orange County, Pittsburgh, and the Prospect School suggests a broader, more important meaning. It is that "educational testing" can mean...
the examination of and experimentation with our educational assumptions, ideas, and practices. In this sense, these four locations have a lot in common. The technology used is almost irrelevant, be it Rasch-scaling, new techniques for testing reading comprehension as in the DRP, monitoring techniques, or tools for documentation. What is common is “educational testing” as a human attitude, a willingness to critically examine educational practice, not just with test scores—numbers—but with other people involved in teaching and learning. Indeed one of the most striking features of conversations with people in these four places was their willingness to engage in what the Prospect School would call reflective conversations about what they were doing. It reminds me of the closing passage in a book by Stanley Joel Reiser (1979), which traced the history of technological innovation and professional practice in the medical field. Despite the limits of the analogy between the professions of medicine and education, Reiser’s comments nevertheless seem appropriate:

If physicians in general come to accept a fundamentally mechanical view of human beings, in a world that is more and more enlamed 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 a human being and regain and reassert his faith in his own medical judgment.

“One need not be a dyed-in-the-wool social Darwinist to recognize that standardized testing is increasing because it serves some important social functions.”

This is not a bad message with regard to improving educational testing. There may be some new and promising testing technologies on the scene. Test scores may be going up. But don’t put much stock in any one testing technology. It is much more important to engage the judgment and concern of people directly involved—students, teachers, and parents—in testing ideas and expectations about teaching and learning.

References


Walt Hanev is Senior Research Associate, Boston College Center for Teaching, McGuinn 531, Chestnut Hill, Massachusetts 02167.