aft A Union of Professionals

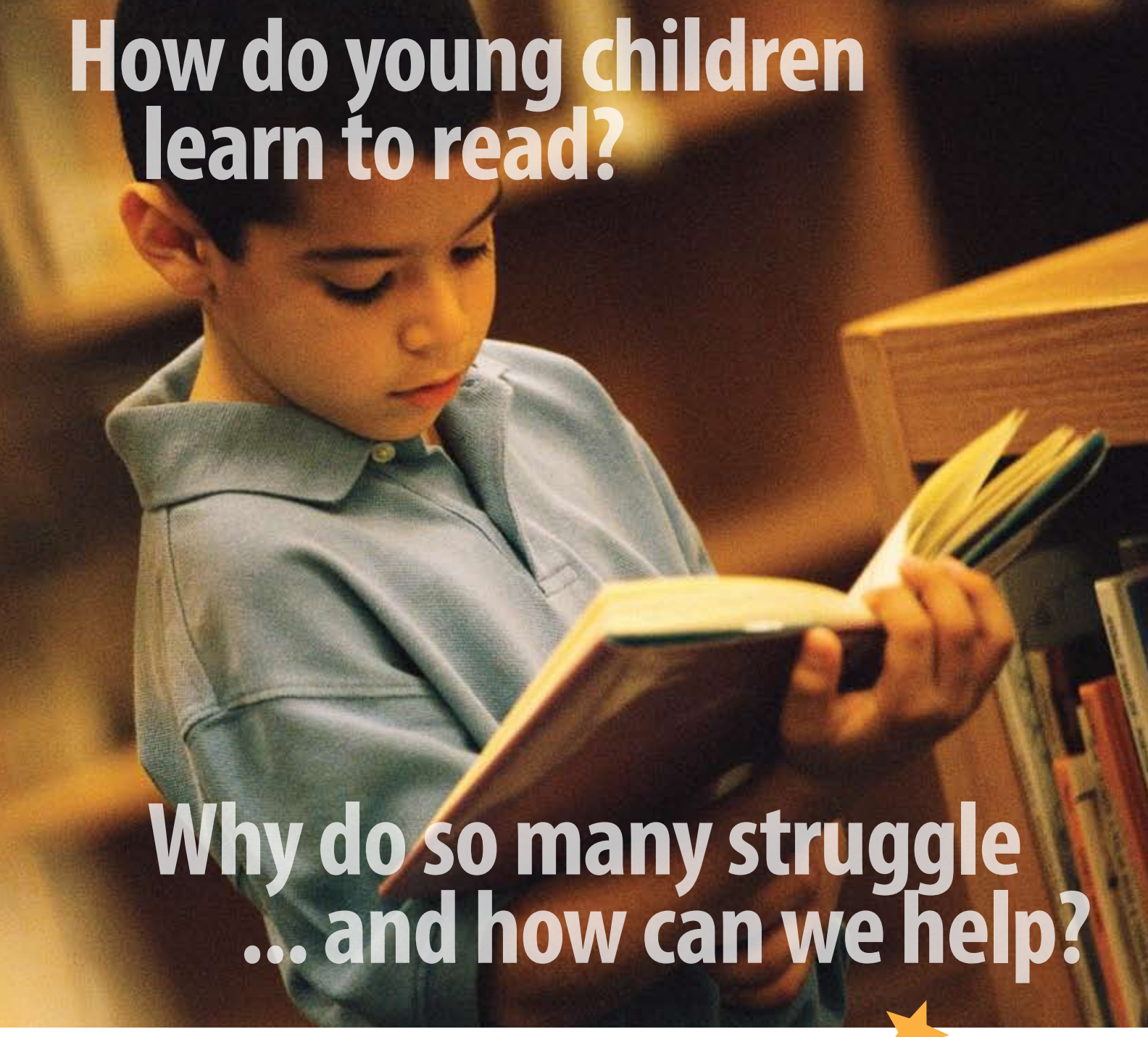
VOL. 32, NO. 3 | FALL 2008

AMERICAN Educator

A QUARTERLY JOURNAL OF EDUCATIONAL RESEARCH AND IDEAS

In Our Hands

WHY TEACHERS SHOULD GUIDE AND GUARD THE TEACHING PROFESSION



READING rockets
www.ReadingRockets.org

For answers, look to Reading Rockets.org
a multimedia literacy website for parents & educators of kids in preschool through grade 3. Visit us online to watch Toddling Toward Reading, our new PBS show!

15% discount at LearningStore
AFT members only! Code: AFT15
www.learningstore.org

Reading Rockets is a national project of public television station WETA in Washington D.C., and is funded by the Department of Education, Office of Special Education Programs

AMERICAN Educator
VOL. 32, NO. 3 | FALL 2008

Cover Illustration by James Yang

AMERICAN EDUCATOR (USPS 008-462) is published quarterly by the American Federation of Teachers, 555 New Jersey Ave. N.W. Washington, DC 20001-2079. Phone: 202/879-4400
www.aft.org

Letters to the editor may be sent to the address above or to **amered@aft.org**.

AMERICAN EDUCATOR cannot assume responsibility for unsolicited manuscripts.

Please allow a minimum of four weeks for copyright permission requests.

Signed articles do not necessarily represent the viewpoints or policies of the AFT.

Although advertisements are screened as carefully as possible, acceptance of an advertisement does not imply AFT endorsement of the product or service.

AMERICAN EDUCATOR is mailed to all AFT teacher and higher education members as a benefit of membership. Non-AFT members may subscribe by mailing $10 per year by check or money order to the address on the left.

MEMBERS: To change your address or subscription, notify your local union treasurer.

Periodicals postage paid at Washington, D.C., and additional mailing offices.

POSTMASTER: Send address changes to AMERICAN EDUCATOR
555 New Jersey Ave. N.W.
Washington, DC 20001-2079.

Advertising Sales Representative
Karen Dorne Media Sales
319 Harrison Ave.
Westfield, NJ 07090
Phone: 908/233-6075
Fax: 908/233-6081

Laying a Better Foundation to Teach Elementary School Math

WHAT DO ELEMENTARY SCHOOL teachers need to know to teach mathematics? Are teacher preparation programs delivering it? These are the key questions behind a new report by the National Council on Teacher Quality (NCTQ), *No Common Denominator: The Preparation of Elementary Teachers in Mathematics by America's Education Schools*.

Since there's no well-developed body of research on the mathematics preparation that aspiring elementary teachers should have, NCTQ spent two years working with its own mathematics advisory group, as well as with a variety of math educators, mathematicians, social science researchers, mathematics associations, and ministries of education in other nations, to develop a set of teacher-training standards. It then used those standards to evaluate elementary education programs at 77 higher education institutions.

Drawing primarily on an analysis of course syllabi and textbooks, NCTQ found that only 10 of the programs, or 13 percent, offered adequate content courses—and 5 of those 10 still could not be wholeheartedly endorsed because they fell short on math methods coursework. Only the University of Georgia was noted for exemplary teacher preparation.

One of the principal findings, highlighted in the table below, was that most education schools fail to devote sufficient time to teaching the four areas of mathematics that are critical for elementary teachers to understand.

NCTQ also found that two-thirds of courses studied either use no mathematics textbook or use a textbook that is inadequate in one or more of the four critical areas. The report states that "the algebra portions of the textbooks are the weakest, with the majority of textbooks earning scores low enough to label them unacceptable for use in algebra instruction." In addition, not one school offers an exit test that establishes whether prospective elementary teachers are prepared to teach mathematics.

To address these problems, NCTQ recommends, among other things, that education schools require three courses that cover elementary and middle grades math content (including algebra, which is taught in the middle grades in many countries), as well as one math methods course that emphasizes numbers and operations. It also calls for the development of a textbook with both content and methods: "This ideal 'combo-text' would augment a core of solid mathematics content with discussion of a process for continuous improvement of instruction focused on student learning."

The entire report, along with a sample mathematics test that NCTQ says every prospective and practicing elementary teacher should be able to complete without a calculator, can be found at **www.nctq.org/p/publications/reports.jsp**.

Deficiencies in Mathematics Instruction for Teachers

Critical areas	Recommended distribution (hours)	Estimated mean of courses in sample (hours)
Numbers and operations	40	27
Algebra	30	4
Geometry and measurement	35	21
Data analysis and probability	10	9

With Census in Schools, Students Can Ask: How Many Toy Stores Are in My State?

CREATED BY THE U.S. Census Bureau, Census in Schools is a program that incorporates census data, such as housing, economic, and geographic information, into free lesson plans and classroom activities. The lesson plans range from teaching students in grades K-2 how to read a map key, to teaching vocabulary such as reapportionment and gerrymandering to 11th and 12th graders.

The program's Web site features a colorful map (shown right) for elementary students that links to state information, including the capital, population data, and even the number of toy stores.

Visit Census in Schools online at **www.census.gov/dmd/www/teachers.html**.

In Our Hands

Teachers Should Guide and Guard the Teaching Profession

What are the hallmarks of a profession? Formal qualifications, a shared code of conduct, specialized knowledge—these and many other qualities are all important, but there's one that teachers should carefully consider: responsibility not just for the quality of your own work, but for that of your peers.

Doctors have their medical boards and attorneys have their bar associations, but most teachers have no such opportunities to take responsibility for their profession. Advocates of peer assistance and review (PAR), a program that gives teachers the lead in guiding and guarding the teaching profession, want that to change. Like doctors and lawyers, shouldn't teachers set the standards for their own profession, help newcomers meet those standards, offer intensive assistance to anyone who is struggling, and recommend the removal of those individuals who, after receiving assistance, are not meeting those standards? Are any of these things really better left to administrators?

Members of the American Federation of Teachers are clearly leaning toward taking greater control of their profession. Earlier this year, a poll of the AFT's teachers found overwhelming support for the idea of having experienced, specially trained teachers mentor and evaluate new teachers—72 percent said their reaction was either very or somewhat positive, and just 8 percent said their reaction was very or somewhat negative. No doubt, that's why the resolution on peer assistance and review, which offers support to locals interested in adopting a PAR program for new teachers, passed so easily at the AFT's 2008 convention.* The AFT's poll also found strong support for assisting and evaluating tenured teachers who are struggling—58 percent said their reaction was either very or somewhat positive, and just 21 percent said very or somewhat negative.

Whether your reaction is positive or negative, learning more is worthwhile. In the following article, Jennifer Goldstein, who did a multiyear study of a peer assistance and review program in California, offers an in-depth comparison of traditional teacher evaluation and PAR. Then, on page 12, Dal Lawrence (who created PAR through collective bargaining while president of the Toledo Federation of Teachers) and two teachers (who have firsthand experience with PAR) talk about what PAR means for professionalism and how combining assistance and evaluation—when done right—can make each more meaningful and powerful.

There are right and wrong ways to address teacher evaluation. Unfortunately, some policymakers and administrators across this country are ready to toss out both traditional, principal-driven teacher evaluation and peer assistance and review. What's their alternative? Complex statistical models that rank teachers according to their "value added." Such models reduce teaching to nothing more than gains in students' test scores. And, as if that weren't bad enough, the models are far, far from perfect. Starting on page 18, Harvard University Professor Daniel Koretz discusses the benefits and limitations of value-added models, explaining that although they do offer some useful information, they should not be used to make any high-stakes decisions. In Koretz's words, "Value-added-based rankings of teachers are highly error-prone."

Once you understand the technical problems with these models, it's clear that value added cannot and should not replace a thorough, thoughtful evaluation of teacher performance. And, once you grasp the many benefits of frequent, ongoing, and interdependent assistance and evaluation, it's clear that traditional, principal-driven teacher evaluation is no match for peer assistance and review.

–EDITORS

ILLUSTRATED BY JAMES YANG

* To read the AFT's peer assistance and review resolution, go to **www.aft.org/about/resolutions/2008/peer_assist.htm**.

Taking the Lead

With Peer Assistance and Review, the Teaching Profession Can Be in Teachers' Hands

By Jennifer Goldstein

I started teaching right out of college. I lacked a teaching credential or any preparation for the job, but nonetheless was given full responsibility for a class of fourth graders in Compton, California. As soon as I found out I would be working at Rosecrans Elementary, I jumped in my car and drove the 30 or so minutes to Compton from the Westside of Los Angeles; having interviewed at the district office, I had not yet seen the school itself. It was summer and the campus was deserted, but Major Thomas, the plant manager, humored my enthusiasm and walked me around. He opened an empty classroom and let me be. I stood there alone, taking in the room with tears in my eyes. Empty classrooms have an almost magical quality, a loud silence full of immense possibilities, and that one on that day even more so for its dilapidation, which I saw romantically: bare wood floors, old-fashioned wood and metal desks and chairs, sunlight streaming in through big metal-grated windows. I didn't yet know that elementary classrooms need rugs or carpets, that there would never be enough desks or chairs, or that the windows would be broken into anyway. I stood there at 23 years old the proudest I had ever been in my life: I was going to be a teacher.

I eventually took ownership of Room 9, which became filled with an always fluctuating number of amazing children. Most were second language learners, some spoke no English, and few could read fluently in any language let alone at grade level. All had fabulous stories to tell, and most were eager to learn. But I had absolutely no idea what to do with them. I mostly used the pedagogical tools of randomness and inconsistency, punctuated with frustrated yelling. Having no vision of a big picture, I did my very best day by day to get by, which was in no way satisfactory for kids who genuinely needed me to teach them something.

I was relatively fortunate that first year to teach across the hall from a quite competent veteran teacher, my assigned "buddy." Actually, Charlotte had only been teaching for three years, but that made her a veteran in Compton; more importantly, she was a bit older, had children of her own, and simply had experience and maturity that I lacked. Charlotte saved me from as much as she could that year, given her own teaching responsibilities. I don't recall actually ever *meeting* with Charlotte per se; it was more support on the run. She handed me lessons to implement, took kids with whom I was struggling on occasion, and declared sole responsibility for planning for the bilingual instructional assistant we shared. That instructional assistant spent one hour in my room three times a week that year working with a group of students, and I have not the slightest idea what she did while there. It is just one example of the degree to which the whole year was a blur. In the end, Charlotte never did actually see me teach, nor I her. When the bell rang and the doors closed, I was on my own.

The other person who might have been expected to provide support or otherwise supervise the teaching my students received was, of course, the principal. She made one visit to my classroom that year, an occasion that stands out amidst the blur. On April 15, the day teacher evaluations were due at the district office, she came in during a lesson, asked me to sign a form, and promised me I would like what it said. I was thus initiated to the quality-control mechanism of my profession.

* * *

You have likely heard some version of this story many times, but its need for attention has become no less urgent. Like so many

Jennifer Goldstein is a faculty member at the Baruch College School of Public Affairs, City University of New York. Her work focuses on teacher workforce quality, teacher professionalization, and the distribution of leadership in urban schools and districts. She was previously an elementary school teacher in Compton and Campbell, California. *This article is adapted from two main sources. The introduction is drawn from Goldstein's forthcoming book on peer assistance and review, due out in 2009 (© Teachers College Press, Teachers College, Columbia University). The remainder of the article is drawn primarily from "Easy to Dance To: Solving the Problems of Teacher Evaluation with Peer Assistance and Review," published in the* American Journal of Education *113 (May 2007) © 2007 by The University of Chicago.*

ILLUSTRATED BY JAMES YANG

marginalized school districts across the United States, Compton schools serve low-income Latino and African American students. My students were attending the elementary school ranked 24th out of 24 in the district ranked lowest in the state of California at the time. Arguably, these were the students most in need of a high-quality teacher. Yet I was unprepared and uncertified to teach. I was in an organizational system designed neither to improve my performance nor assess it. In addition, after three years—or right around the time research predicts my teaching performance would improve significantly[1]—I left the district.

In school districts like Compton all over the country, there are simply not enough qualified teachers willing to staff classrooms.[2] As a result, administrators hire teachers who are uncredentialed or credentialed in a different field. In California, for example, 1 in 15 teachers—approximately 20,000 total—were underprepared in 2004–05, and notably 85 percent of these teachers were concentrated in schools serving predominantly students of color.[3] The urgent reality is that 30–50 percent of new teachers in high-poverty schools then leave teaching within their first three to five years on the job, and those without support are 70 percent more likely to leave than those who receive a mentored entrance to teaching.[4]

This article explores one high-leverage policy that administrators such as those in Compton could implement to attract teachers who are qualified, support and guide them to develop into teachers with high-quality performance, and retain them beyond their initial years in the job, while also removing from classrooms those teachers who do not display competency even after having received intensive support and mentoring. The policy is called peer assistance and review (PAR), and it is a model of teacher-based instructional leadership that has the potential to transform teaching practice by transforming teacher evaluation. PAR shifts evaluation from the typical cursory review by a principal with a checklist, to a year-long process that involves both frequent, ongoing, classroom-based assistance and a careful, standards-based review.

PAR (pronounced as the word "par" and also referred to as "peer review") experienced a very specific birth in Toledo, Ohio, in 1981, the result of many years of effort by Dal Lawrence, the then-president of the Toledo Federation of Teachers. (To learn about Lawrence's eight-year struggle to create PAR, and what teachers think of the program, read the interview with Lawrence and two Toledo teachers on page 12.) Over the next two decades, a handful of districts—Cincinnati and Columbus, Ohio; Poway and Mt. Diablo, California; Rochester, New York; Dade County, Florida; and Salt Lake City, Utah—undertook their own versions of the "Toledo Plan" of peer review, still a common blueprint of the policy.*

Broadly speaking, PAR relies on "consulting teachers" (CTs)—teachers identified for excellence who are released from classroom teaching duties for two to three years, usually full time. During that time, the CTs provide mentoring to teachers new to the district or the profession, and intervention support for identified veteran teachers experiencing difficulty.† The consulting teachers also conduct the formal personnel reviews of the new teachers in the program, and in some cases they participate in the reviews of the veteran teachers as well. In some programs, and for my purposes here, teachers in either the new or veteran category are collectively called "participating teachers" (PTs). The duration of participation in PAR is usually one year, although some programs allow longer. During this time PTs have to meet specified quality standards or face removal from the classroom.

Consulting teachers' activities include helping with short- and long-range planning, locating curricular resources, advocating

† Veteran teachers, who make up a relatively small percentage of teachers in most PAR programs, are most typically placed in PAR for intervention upon receiving an unsatisfactory evaluation from the principal, although in some districts other avenues for referral exist. Intervention cases are reviewed for validity at the outset; the shortcomings in the teacher's performance must involve instructional matters, as noninstructional matters are not the purview of the PAR panel. Many PAR programs also include an alternative evaluation option for tenured teachers who are meeting standards.

* To learn more about the Toledo Plan, see **www.tft250.org/the_toledo_plan.htm**.

for the participating teacher with the principal, jointly observing other teachers, and providing general emotional support. The vast majority of CTs' time, however, is focused on observing PTs teaching and providing feedback and suggestions on instructional strategies. Each CT-PT pair works together to create an individual learning plan that focuses their work together.

The consulting teachers report to a districtwide joint teacher/administrator board called the "PAR panel."* The panel is typically co-chaired by the union president and the director of human resources (or some other high-ranking district administrator), and has a combination of teacher and administrator members. The panel holds hearings several times a year, at which consulting teachers provide reports about participating teachers' progress, present any concerns about teaching performance, and receive suggestions for improving their work with PTs. Depending on the locally agreed-upon details of the program, the consulting teachers may make recommendations about the continued employment of each participating teacher at a spring hearing, and sometimes sooner. After listening to and questioning the consulting teachers, the panel makes its employment recommendation, most typically to the superintendent of schools, who then makes a recommendation to the school board, the ultimate arbiter of personnel decisions.

The vast majority of consulting teachers' time is focused on observing participating teachers and providing suggestions on instruction. Each pair of consulting and participating teachers creates an individual learning plan that focuses their work together.

PAR in Rosemont: An Effective Model of Teacher-Based Instructional Leadership

Almost 10 years ago, a new law in California gave me the opportunity to look closely at the PAR model of teacher-based instructional leadership. In 1999, California Assembly Bill IX marked the first time PAR was instituted statewide and the first time a major district had implemented the policy in over a decade. By 2002, a state budget crisis and competing state legislation had begun to chip away significantly at California's PAR programs. I conducted a study of PAR in that window of time (primarily between 2000 and 2002) in one urban district in California, which I have given the pseudonym Rosemont. The study involved a year of full-time fieldwork and data that span four years, and is among the most in-depth investigations of a PAR program to date.

Under the California law, most PAR program details were left to local school districts. As a result, and like PAR programs nationally, California districts created programs that often looked different from one another: many did not include new teachers in their PAR programs, as the state law required the program only for veteran teachers who had received an unsatisfactory evaluation from their administrator; many did not create full-time positions for consulting teachers; and many did not involve consulting teachers in personnel reviews. In Rosemont, however, teachers—via both the consulting-teacher and PAR panel-member roles—were given substantive authority in the development of teaching quality, as well as in deliberations about employment for both new and veteran participating teachers. I do not claim that Rosemont's results are necessarily representative of all efforts called "peer review," but rather that they demonstrate what is possible when union leaders and district administrators bring a fresh perspective and ample resources to assisting and reviewing teachers.

For the first year of the PAR program, Rosemont selected 10 consulting teachers, who supported 88 beginning teachers and 3 veteran teachers across 28 schools. All consulting teachers and panel members participated in the study, which included repeated observations, interviews, and surveys. In addition, 16 principals and 57 participating teachers completed surveys, and I conducted interviews with 3 key district administrators, 11 principals, and 15 beginning teachers. (I did not interview any of the veteran teachers due to the sensitivity of their situations.) Given the small number of veteran teachers in the program, this article focuses on the consulting teachers' work with new teachers, providing an overview of the major components of PAR and how it differs from traditional teacher evaluation by a principal.

My examination of PAR in Rosemont yielded six key features that distinguished it from teacher evaluation as typically conducted by principals: (1) the amount of time spent on evaluation, where consulting teachers assisted and reviewed a caseload of participating teachers full time; (2) the relationship that professional development has to evaluation, where reviews were linked to assistance, including matching consulting and participating teachers by grade and subject, and using performance standards; (3) the transparency of the evaluation process, where PAR panel hearings and consulting teacher meetings made teachers' prac-

* Note that different districts use different terms for similar program roles and components. For example, CTs are sometimes called coaches, and the panel is sometimes called a governing board. Participating teachers are sometimes referred to as interns (if a beginning teacher) and intervention cases (if a veteran). Regardless of the terms used, these core components remain essentially the same.

tice and evaluative decisions about that practice more transparent; (4) the nature of labor relations, where the teachers' union was part of the process; (5) the level of confidence in the decision-making process, where the PAR process seemed to generate more confident evaluative decisions; and, ultimately, (6) the degree of accountability, where consulting teachers were willing, when necessary, to recommend nonrenewal and panel members, after ensuring that sufficient evidence had been presented, were also willing to recommend nonrenewal.

While taking a closer look at each of these six distinguishing features, this article builds on the literature that demonstrates the flaws of traditional teacher evaluation, and it posits that the more professional model of PAR shows promise. For each of the six features, I first draw on existing literature (and sometimes data from Rosemont) to highlight the problems with traditional teacher evaluation. Then, drawing primarily on data from Rosemont and occasionally on other studies of PAR, I present the ways that PAR can address these problems.

1. Making Time for Assistance and Review

Problem: Principals are overwhelmed by the demands and expectations currently placed on them,[5] with little time for instructional leadership at a time when the focus on accountability for instructional results has increased.

Lack of time affects both the assistance and review of teachers. In Rosemont, for example, principals admitted that they cut corners with their evaluations, by necessity. Principals described the "wiggle room" or need to be "creative" in doing their evaluations—typically doing fewer than desired, or even required, on teachers perceived to be performing acceptably. One principal noted simply that "the current evaluation process really is a sham, it's a joke." Many principals identified their need to be in classrooms and know what is going on across the school but described merely popping their heads in and out. Or, as one principal admitted, some saw teachers based on the whims of geography: "It probably depends how close they are to my office, too. Things as dumb as that even, whether they're on my trip. Like I'm going to go to the cafeteria in a few minutes and if they're on the way up, I'll probably see them more often than if they're over in the corner somewhere."

With the traditional evaluation process used in Rosemont, principals, as well as consulting teachers and panel members, agreed that principals' lack of time allowed teachers not meeting standards to slip through the cracks. It also invariably allowed some of those who could have developed into excellent teachers to slip through the cracks as well, through attrition, since teachers who are not systematically supported are far more likely to leave the profession.

Solution: The consulting teachers were released from classroom teaching responsibilities and focused on their participating teacher caseloads full time. By contract, consulting teacher caseloads were 12-15 participating teachers. In reality, because consulting teachers were involved in program development in the first year of implementation, they carried caseloads of approximately 10 participating teachers that year. In addition, consulting teachers developed a formula whereby assisting a veteran teacher counted as two new teachers when constructing caseloads, given what they perceived as the larger emotional drain and investment of time needed when working with a veteran teacher.

All consulting teachers were expected to visit their participating teachers an average of one time per week, to make some unannounced visits, and to conduct three formal observation cycles during the year for personnel review purposes, presenting one at each panel hearing. Participating teachers did report meeting with their consulting teachers on average once per week, especially at the start of the school year, but this ranged from "at least once a week" to once every two to three weeks, as consulting teachers' visits to participating teachers' classrooms typically became less frequent for more effective PTs as the year progressed. Some consulting teachers preferred to come by informally and unannounced, while others had a set time to visit every week. Noted one participating teacher:

> On Tuesday, we had a pretty routine schedule, which made it a lot nicer. I knew she was coming during second and third period every Tuesday, so I could count on that, I could make questions ahead of time that I knew I was going to want to ask. I'd teach during second [period]. So, she would typically observe during that time, and almost every time, she would give me written feedback on things that looked good and ideas for improvement. And then, third period's my prep, so we could talk then.

Participating teachers reported that CTs made their ongoing accessibility clear at the beginning of the year, provided e-mail addresses and cell phone numbers, and could be reached as needed. Forty-seven percent of participating teachers and 80 percent of principals interviewed initiated comments on the availability of the CTs and the amount of time they were able to spend working directly with PTs. The structure of CTs' full-time release from classroom teaching responsibilities allowed them to be on call to meet PT needs as they arose. Noted one consulting teacher, "There were a number of times where teachers called me on just specific little issues, whether it was a parent issue, a child abuse issue, an issue having to do with their principals—just little things, how-tos, that were very simple to solve, but having that relationship was important."

Overall, consulting teachers' time allowed a high level of involvement in the details of participating teachers' day-to-day lives that principals simply could not match, as they were busy running schools. A principal contrasted what she could provide to beginning teachers with what the CT provided: "Before PAR started I had Friday meetings with my new teachers and they would go forever, because they'd have a million questions and I would answer them and I would write down things that they needed and I would try to support them. But I can't model a lesson in every one of their classrooms, and I can't do the kinds of things that a PAR consulting teacher can do, because I'm running the whole school." The participating teachers recognized the difference between what their CT could give them versus what their principal could give them. Two of the 15 participating teachers interviewed had had negative experiences with their principals and therefore were especially grateful to be involved in PAR. The majority of PTs, however, regarded their principals with respect for their seemingly insurmountable jobs, and simply viewed the PAR program as a logical way for them to get

desperately needed assistance. One PT made this compassionate contrast: "My consulting teacher is a really good listener. I think more than my principal, my CT is a deeper, more thoughtful listener. She is doing something very specific for me, where my principal is doing a million things for everybody.... My principal wants to give me his attention, he's trying, ... but no one can do everything."

Data from established PAR programs suggest a positive effect on the retention of new teachers. In Columbus, 80 percent of new teachers are still in the district after five years, and in Rochester, the retention rate was 85 percent for the five-year period from 1998 to 2003.[6] In Rosemont, principals reported being able to relax a bit about their new teachers with the implementation of PAR, knowing the teachers were getting consistent assistance and review. Survey data indicated that principals, panel members, and consulting teachers all thought PAR had a positive impact on principals' ability to do their jobs well.

2. Linking Professional Development and Evaluation

Problem: Teacher evaluation has generally been defined as a mechanism for appraisal in order to determine fitness for employment rather than a means for improving performance. Typically, the principal's evaluation is completely separate from any professional development a teacher may receive from formal or informal support providers. Key here is that very often administrators conducting traditional evaluations are not privy to the knowledge and perspective that these support providers have about a given teacher's performance. As a result, principals' evaluations are often based on very little data,[7] limited to infrequent formal classroom observations that are almost always announced and may be quite short in duration.

Compounding the problem, principals have to evaluate all of the teachers in the school, and therefore often lack expertise in the specific content or grade level of many of the teachers for whom they are responsible.[8] In addition, principals are often not well trained to conduct the evaluations.[9] Such a system is not about learning or developing as a professional, but is merely the proverbial hoop through which to jump.

Solution: As a result of consulting teachers' full-time focus on the professional development of participating teachers, PAR evaluations were based on ongoing observations throughout the year and intimate knowledge of a PT's classroom—rather than the notorious "dog and pony show" of most teacher evaluation systems. Linking review to assistance through PAR (a) built trust and rapport; (b) provided participating teachers with ongoing instructional feedback; (c) created individualized assistance; and (d) grounded the reviews, and the training of the CTs as reviewers, in performance standards for teaching.

a. Trust: Most consulting teachers felt that supporting participating teachers' day-to-day needs, especially at the beginning, helped develop rapport and build trust. While strong mentor programs often focus on trying to move mentor-mentee interaction beyond emotional support to substantive dialogue about teaching and learning, the reality remains that new teachers often do need emotional support.[10] For some PTs, the trust needed to speak openly about teaching and learning was developed by first knowing the CT was there to help. Noted one PT: "I think one benefit is just knowing that there is someone out there that is on your side, who you can go to to talk things through." In contrast to a fear sometimes raised about PAR, linking assistance and review did not appear to have a deleterious effect on PTs' trust in their CTs in most cases.[11] (For more on how consulting and participating teachers build their relationship, see the interview with two Toledo teachers on page 12.)

b. Ongoing feedback: In addition to building trust and rapport, however, the heart of the PAR program was ongoing feedback to participating teachers about how to teach. Wherever possible, PAR consulting teachers were paired with PTs by grade and subject matter. For several PTs, this matching was critical to their ability to work meaningfully with their CTs. Noted one: "The difference between my principal and [my CT] is that my CT has experience in biology, and just in sciences in general; she was able to bring materials and suggestions to the class. The principal doesn't have that experience, her area isn't in sciences. My CT would make suggestions about how to go about teaching things, and it would trigger ideas and thoughts for me."

c. Individualized assistance: This grade and subject matching, together with the time consulting teachers had available to work with participating teachers, created an environment of individualized assistance, which CTs often compared to a good teacher's ability to individualize instruction for students. The participating teachers noted that CTs had a high level of familiarity with day-to-day operations in their classrooms, allowing them to provide tailor-made support, such as bringing curricular materials that fit right in with a unit the PT was planning, being able to talk specifically about struggles with certain students, or recognizing when the PT was getting burned out and needed a break. The individualized assistance provided to each PT contributed to informed evaluative judgments. One participating teacher commented, "[My CT] really picked out some things that she thought that I could improve on, something with teaching style and something with assessment. And throughout the year, she really helped me with those things. So by the time she would do a formal evaluation, she could show how I'd improved in those things."

d. Performance standards: Strong evaluation systems include established standards for performance, evaluation rubrics based on those standards, and evaluator training for interrater reliability,[12] although many teacher evaluation systems nationally lack these components.[13] While consulting teachers were not experts in performance standards for teaching at the time they were hired, they poured many professional development hours into becoming experts, and then into becoming calibrated among themselves in using a rubric to evaluate their participating teachers. Participating teachers were evaluated on a slightly modified version of the California Standards for the Teaching Profession, which served as a benchmark throughout the year. Conversations between consulting and participating teachers about instruction were often grounded in standards language.

The consulting teachers' ability to demonstrate participating teachers' growth, or lack of growth, at panel hearings was dependent on solid standards-based assessment documentation gathered over time through ongoing observations. For example, in one case, a principal had hired an uncredentialed teacher one

week prior to the start of school, but quickly concluded that she was not meeting standards. While the consulting teacher was initially skeptical of the participating teacher's chances for success, she was persuaded by the progress the PT was able to make and defended the PT's renewed employment in the district. The consulting teacher became the mediator, translating the principal's broad concerns into concrete specifics on which the PT might improve. Ultimately, the consulting teacher diffused the principal's criticism of the participating teacher at the panel hearing by demonstrating her growth on the teaching standards. The principal's complaints seemed vague and unsupported by comparison. The participating teacher was renewed for employment in the district and placed at another school. This fluency in standards language gave the CTs legitimacy with both principals and panel members, as well as with PTs. Several principals were so impressed with the standards-based reviews that they asked a CT to teach them the process. Principals and panel members perceived CTs to have developed valuable expertise, which was crucial to PAR's success.

Wherever possible, consulting and participating teachers were paired by grade and subject matter. For several participating teachers, this matching was critical to their ability to work meaningfully with their consulting teachers.

3. Increasing the Transparency of the Evaluation Process

Problem: Teaching has been a notoriously isolated occupation, with individual teachers behind closed doors with their particular group of students,[14] and occupational norms that typically prevent teachers from "intruding" on one another's practice.[15] Noted one Rosemont principal, "The 11th Commandment is you don't speak ill of another teacher. I taught for seven years next to this nice person, just an awful teacher, and I could hear her through the wall, hear the kids and stuff and I would go over and have to quiet them down, just to kind of bring some sanity to it. But it was like the elephant in the living room. Nobody would talk about how awful she was."

Just as teachers work mostly in isolation, so do principals. As a result, another "elephant in the living room" is the small amount of information and input on which most principals base their evaluations. We've already noted that principals' evaluations are typically separated from any information that support providers may have. Another problem is that, alone with their observation notes or checklists, principals then typically make evaluation decisions on their own, not needing to defend their decisions to another colleague, let alone a panel of colleagues. Research has documented that, historically, principals have given inflated ratings and few negative evaluations for a variety of reasons, including minimal observation data[16] and a potent desire to avoid conflict.[17] This tendency may be understandable, but it does little to ensure a competent teacher for every student.

Solution: PAR provided several ways of avoiding some of the opacity of traditional teacher evaluation. First, consulting teachers worked in multiple schools across the district based on grade and subject matching. In this way, CTs brought a broad, districtwide perspective to assessment, and a CT was not paired with a PT where there was a conflict of interest or other personal connection. (Some smaller districts with PAR programs have formed consortia, pooling consulting teachers across districts in order to accomplish this goal.)

Next, PAR opened the door to practice, altering the historic isolation of teaching by placing a mentor in PTs' classrooms on a frequent basis. While certainly not unique to PAR, the ongoing nature of PT-CT interaction is a critical piece in the quality of the reviews, because increasing the publicness of practice is likely to increase the amount of data on which reviews are based. Noted one PT:

> Had the vice principal come up to do the evaluation, she would have had no idea what it's like on a normal basis, when the vice principal was not sitting in the back of the room. I really like the idea that my CT did my evaluations. Who better than someone who really has seen the whole picture? She had an idea of where I had started, and how much I had grown. She knew the struggles I had had, so she could look to see if I had addressed those. I really liked that there was some kind of benchmark.

Finally, and most importantly, PAR created formal teams of colleagues and a structure for holding evaluators accountable for their work. Given a larger amount of data about a teacher upon which to base both ongoing assistance and review, PAR provided a mechanism whereby multiple educators were in communication with one another about that data. CTs met as a group all day every Friday, and some of this time was spent discussing PT cases and seeking advice from one another. In addition, CTs formed pairs of "critical friends," and occasionally met to discuss their PT cases or visit a PT's classroom together for a second pair of eyes.

Consulting teachers also conferred with principals. CTs were focused on classroom practice, whereas principals had a perspective about the PT as part of the school community. By the

second year of PAR, Rosemont created a format where both the CT and the principal observed a PT and then conferred, in order to be sure they were in agreement regarding professional development needed and/or the recommendation to the panel regarding the PT's renewal status.*

In addition, the most significant and formal examinations of PT practice were the PAR panel hearings that occurred multiple times throughout the year.[19] CTs reported to the panel roughly three times a year on PTs' growth and/or problematic practice, first with extensive documentation and then with oral presentations. The teachers and administrators sitting on the panel offered suggestions regarding support the CTs might try, and held the CTs accountable for providing sufficient assistance in order for the PTs to have the opportunity to improve. In this way, an individual teacher's practice became a district concern. In a few instances, CTs were challenged to provide more evidence for their employment recommendation or even to return to the PT for a few more weeks for one last effort. Noted one CT: "I was tap dancing around giving a decision of nonrenewal, and they asked me directly, 'What is the evidence for keeping this person?' And I really didn't have enough. They held me accountable, and that was appropriate."

The PAR panel held the consulting teachers accountable for providing sufficient assistance for the participating teachers to improve. In this way, teachers' practice became a district concern.

* Dal Lawrence, the former president of the Toledo Federation of Teachers who initiated PAR more than 25 years ago, has argued vehemently that principals should not be involved in the peer review process for legal reasons. His argument is that there needs to be one clear evaluator, otherwise there is a possibility for disagreement that can cause a loss to an unsatisfactory teacher in arbitration.[18]

Bringing Peer Assistance and Review to Your District

Educators interested in implementing peer assistance and review (PAR) should carefully consider the following challenges, gleaned from the study of Rosemont and other efforts.

Ensuring Consulting Teacher Quality

The perceived success of the program appears to be based largely on principals' and panel members' confidence in the consulting teachers (CTs). It follows that CTs should be selected very carefully. Consulting teachers must be regarded as master teachers, and in Rosemont the selection process included classroom observations by two panel members. The consulting teachers were also required to demonstrate prior success mentoring a peer, including a letter of recommendation from a teacher they had mentored. Finally, the consulting teachers had to be above reproach. Given the authority that CTs held with respect to employment recommendations, it was critical that the selection process appear unbiased and without favoritism. Once selected, it was imperative that consulting teachers received training in coaching methods, teaching standards, and assessment, and that they remained vigilant with respect to confidentiality.

Defining Good Teaching

Effective PAR programs require agreed-upon standards of practice and performance rubrics, which form the foundation of the work between participating and consulting teachers. In addition, evaluative decisions must be beyond reproach, with detailed standards-based documentation from the classroom. The challenge in many districts is that educators have not defined quality teaching or made their priorities and values clear—a necessary step for a transparent evaluation process. They also may not find themselves in agreement when they do make their values explicit. Creating these conversations, and owning (rather than importing) the standards of practice that grow out of them, are crucial steps in the PAR process.

Reframing Labor Relations

A critical issue for PAR implementation is the level of trust between teachers and administrators.[1] For this reason, most school districts begin PAR programs with new teachers only, since the idea of apprenticeship is far less controversial among teachers than peer intervention with veterans. The expansion to include intervention cases typically occurs once a program has been in place successfully for a few years. This was not the case in California, where the state legislation specifically targeted veteran teachers. As a result, Rosemont and other districts across the state were required to skip the trust-building phase of PAR.

Reframing Instructional Leadership

Despite their complaints that they do not have time to do evaluations well, administrators are often quick to defend their turf. Principals' hesitancy to relinquish authority for teacher evaluation is understandable and, where it signals professional commitment to teacher quality and instructional leadership, laudable. The problem and its solution lie in the conception of instructional leadership. Rather than define an instructional leader as one who directly provides the instructional support for teaching and learning, with PAR, principals enact instructional leadership by communicating regularly with CTs, meeting with the panel, and conducting the personnel evaluations of those teachers not in PAR.

Building Bridges to Mentoring Programs

Some educators may adhere to the notion that assistance and review must be

The panel's expectations that the consulting teachers' assessments be standards-based stood in contrast to the ubiquitous "I know good teaching when I see it" that has plagued much of traditional teacher evaluation. Teaching standards or "protocols of practice"[20] depersonalize the process, creating a review that focuses on the teaching practice rather than the person. The teachers' union president noted, "We're trying to institute standards for teaching so that people will be playing on a common playing field, with common rules. Hiring and firing decisions would be made centrally. They would be based upon standards rather than the whim of a particular individual." In the union president's eyes, PAR served two purposes: reducing principals' ability to fire new teachers at will and increasing accountability for poorly performing teachers.

As the panel made individual teachers' practice a districtwide concern, it also increased accountability for administrators by identifying "red-flag situations" in schools across the district. For example, the panel identified some cases of principals failing to give beginning teachers a sufficient opportunity to succeed, such as an assignment of four preparation periods across three classrooms on two different floors of a building. Extremely challenging situations like this complicated the CT's job of diagnosing and assessing a PT's practice and potential. In such cases, the associate superintendent on the panel addressed the situation with the principal directly and sometimes required that the conditions for the new teacher be altered.

The panel process was certainly not perfect. One of the main criticisms of the PAR panel by consulting teachers was that they did not play a critical enough role. For the most part, this seemed to be an issue of time. Hearings typically ran all day for two days, yet most of those involved tended to feel the process was rushed, not allowing sufficient time to go into the depth they would have liked. It is therefore not surprising that some CTs reported feeling that the panel was a rubber stamp on their decision about a PT. While the data reveal increased transparency, there was still plenty of room for growth toward more meaningful involvement of the panel in the process.

4. Involving the Teachers' Union

Problem: The typically confrontational nature of education's labor relations can make the rare attempt at dismissal prohibitively costly and time consuming.[21] Many principals have viewed the union as an unbeatable adversary and often do not try to fire a teacher.[22] Instead, they engage in escape hatches,[23] such as transfers (voluntary and involuntary), resignation, and retirement.[24] One Rosemont principal explained that, with traditional teacher evaluation, "someone allowed me, not correctly, but

(Continued on page 36)

separate in order to ensure trust between mentor and mentee. While that concern was not supported by this research, those interested in implementing PAR must attend to it or face resistance. Rosemont's PAR program benefited greatly by resting on a decade and a half of mentoring efforts in California. Rosemont's consulting teachers were able to enter an already existing statewide conversation about performance standards for teaching and effective coaching strategies, and some of them had already served as mentors in the statewide Beginning Teacher Support and Assessment program. Strong PAR programs require deep knowledge about teaching and learning. If PAR is being considered in a district that already has a mentoring program, it's important to work with the current mentors so that the PAR program benefits from their knowledge and so that the mentors have an opportunity to consider the benefits that can arise from combining assistance and review.

Paying for PAR

The main cost involved with PAR is the replacement cost of consulting teachers who leave the classroom, which in Rosemont was covered by funding from the state per the state legislation. Other, more minor costs include stipends for teachers on the PAR panel and release days for participating teachers to observe other teachers. PAR programs may result in savings, however, as they weed out weak teachers while they are probationary, avoiding the expense of termination later after they become tenured. Evidence indicates that PAR programs may also improve retention, avoiding the expense of recruiting, hiring, and orienting yet more new teachers. These cost savings are hard to measure; nonetheless, future research should attempt to do so.[2] Two ways that districts may be tempted to try to reduce the cost of implementing PAR are reducing CTs' release time to something less than full time or increasing CTs' caseloads. Either of these approaches risks undermining the effectiveness of the program, since CTs' work with participating teachers hinges on time, and administrators' impressions of the program hinge on the perceived effectiveness of the CTs.

Overcoming the Norms of 100 Years of Bureaucracy

Despite the largely positive response to PAR in Rosemont, it is very difficult to shift norms in the way required by this policy. Principals and panel members had great confidence in consulting teachers' abilities and believed that they were conducting high-quality reviews. But most people—principals, panel members, and CTs themselves—wanted principals to be more involved in the process. Suffice it to say that policymakers and practitioners should be clear about their intentions regarding instructional leadership and responsibility for teacher evaluation when implementing PAR, as people will tend to regress to that which is familiar, namely, principal control. Historically, districts move quickly to blunt the effects of new teacher leadership policies.[3] Educators—whether union leaders, teachers, or administrators—must overcome long-standing norms if they are going to put collective responsibility for professional standards in teachers' hands.

–J.G.

Endnotes

1. A. Urbanski (presentation at the annual meeting of the Teacher Union Reform Network, Santa Cruz, CA, November 1999).

2. L. Kaboolian and P. Sutherland, "Evaluation of Toledo Public School District Peer Assistance and Review Plan" (unpublished report, John F. Kennedy School of Government, Harvard University, Cambridge, MA, 2005).

3. J. W. Little, "The Mentor Phenomenon and the Social Organization of Teaching," *Review of Research in Education* 16 (1990): 297–351; and P. B. Sebring, S. Hallman, and M. Smylie, "When Distributed Leadership Is Called Back" (paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 2003).

Peer Assistance and Review

A View from the Inside

To better understand peer assistance and review (PAR), American Educator*'s editors spoke with three people who know it inside and out: PAR's founder, a consulting teacher, and a participating teacher. Dal Lawrence, former president of the Toledo Federation of Teachers and of the Toledo Area AFL-CIO Council, provides a glimpse of his struggle to create a teacher induction program, and why he thinks the end result, PAR, is so important for teacher professionalism. Audrey Fox, a consulting teacher, and Melissa Joseph, 1 of 10 participating teachers who worked with Fox last year in Toledo, Ohio, discuss their relationship and why they believe the PAR process works well. Fox, who's in her 12th year of teaching and 3rd year (of a three-year term) as a consulting teacher, has mostly taught English at the middle school level. Joseph, who taught for two years in Michigan as a long-term substitute before coming to Toledo, teaches English at Scott High School.*
–EDITORS

Editors: Why is it important for the union to promote teacher professionalism and how does PAR contribute to it?

Dal Lawrence: PAR helps us look at our culture as teachers. Teaching is too often an isolated experience in which teachers take great pride in their classroom, but if they have a colleague down the hall who is having trouble, they typically don't think that's their responsibility. It's the responsibility of somebody in the office. PAR begins to change that concept of responsibility, spreading it throughout the teaching staff.

By almost everyone's judgment, the evaluation of teachers in public schools is broken. Principals are busy people, so they tend to avoid dealing with instructional problems. With PAR, a joint union-management panel accepts responsibility for competent instruction. With intensive peer assistance and a thorough evaluation, you find out who should teach, and you shorten the learning curve for new teachers from about five years to two semesters. The importance of helping new teachers improve was impressed upon me when I started to teach. I had a master's degree in history and six weeks of student teaching. It was at least five years before I was really in command of my ability to teach kids. And I was frustrated most of that time. I had a two-year probationary period, as most teachers in Ohio do, and I had four satisfactory evaluations—even though no one ever appeared in my room.

Audrey Fox: I take pride in my career as an educator, and PAR allows me to feel valued as a professional because, as a consulting teacher, I have to uphold high standards—and I'm also held to high standards. In PAR, we have rubrics for classroom management, teaching procedures, and professionalism. In each rubric there are specific, detailed objectives and descriptions of what a satisfactory teacher looks and sounds like, and what an unsatisfactory teacher looks and sounds like. This allows the communication between the participating teacher and the consulting teacher to be consistent and based on clear standards, not opinions.

When I stand before the PAR panel, I am held extremely accountable. If I have a participating teacher who is unsatisfactory who I am recommending for nonrenewal, I am thoroughly questioned. But I am just as thoroughly questioned for the teacher that I'm saying is satisfactory. I give a very detailed description with specific examples from the classroom. Afterward, the panel members ask me a plethora of questions, seeking more examples and thorough explanations. The process ensures objectivity, thanks to the specific standards and guidelines we are all held to.

Editors: When the union first began advocating for PAR, was there any resistance among teachers or administrators?

Dal Lawrence: We didn't have resistance from teachers. We poll our members every three years, and they have been consistently and overwhelmingly in favor of PAR. Our membership actually supported the idea as far back as 1973. The reason for that is we were asking teachers the right questions, such as, "What do you want to be that you're not now?" They all wanted to be part of a profession respected for its excellence. We looked to the medical model, with its internship and residency, and used it in creating our PAR proposal, which was essentially an induction process for new teachers.

We had resistance from principals. It took us eight years to get PAR adopted. It was finally implemented in 1981. From 1973 to 1981, we were talking to school administrators across the bargaining table and they were saying that we couldn't do this—that it was their job. Then, in 1978, we had a really tough strike. We won it big time. We ended up with a new superintendent and, for the first time, an attorney who was the board's negotiator. In March of 1981, I put the proposal for a new teacher induction process on the table again. The attorney asked why management didn't want to implement it. I said that it's a turf issue. He asked, "We don't fire anybody for incompetence, do we?" I said no. I had looked over the school board minutes for the past five years, and we hadn't fired a single person for incompetence. The next time we met, he again asked how this would work. He pointed across the table at me and said, "We're going to do this, but you're going to have to do something for us." He said if we could use these "expert teachers"—that was his term—to work with senior teachers, including those with tenure, who have severe problems, "you've got a deal." I stuck my hand across the table and that's the way it started.

Editors: Since you conceived of PAR as a program for new teachers, how did you incorporate management's demand for an intervention component for tenured teachers?

Dal Lawrence: We worked out somewhat different procedures for new and tenured teachers. There are two critical differences. One is that when a consulting teacher is working with a tenured teacher, the consulting teacher writes a detailed report for management, the union, and the teacher, but that report merely explains what has happened, it is not an evaluation. The other key difference is that with tenured teachers, the union has to ensure that a fair process is in place such that the tenured teacher's due process rights are respected and that the union upholds its duty of fair representation. In Toledo, we have an attorney who represents both union and management who is called in to review the situation before assistance even begins with a tenured teacher. It's an upfront piece of due process that ensures all procedures are followed.

When I work with school districts interested in implementing PAR, I recommend that the assistance for tenured teachers be a choice for the member who's having trouble. That

member can either face dismissal by management, and the union can provide representation in the traditional way, or that member can say, "Wait a minute, maybe I can get back to meeting standards if I have some help." At that point you can assign a consulting teacher to give the assistance that's needed. After all, if you have a member who is having severe problems, why wouldn't it be the responsibility of a union of professionals to at least offer some help?

Editors: Why is it important to combine assistance and review for new teachers? Why not create a mentoring program?

Dal Lawrence: If we really are professionals, then we ought to accept responsibility for instructional competence. In PAR, practically all the work is mentoring. The evaluation is the summary of the work that the consulting teacher has done with a participating teacher. After that consulting teacher has spent hours working with an individual, the evaluation is not only an evaluation of the participating teacher, it's an evaluation of that consulting teacher's own mentoring. If a teacher fails to meet standards, it's not only that teacher's failure, it's our failure, too. We don't give up easily.

Melissa Joseph: Combining assistance and review makes for a better evaluation because the consulting teacher gets to know you. She's there to work with you, so she knows your strengths and weaknesses better than the principal would. It helps to have someone who is consistent, who is there on your good and bad days, and who is there to help you arrive at a goal—not a one-time pop into the classroom that might happen when a lesson plan isn't going as well as you had hoped.

Audrey Fox: From my perspective as a consulting teacher, I think one of the great benefits of assistance and review being combined is that I have a vested interest in each of my participating teachers. My job is to provide the assistance necessary to take them to successful completion of the program, to be able to say that they're satisfactory in all areas. I'm held accountable for that.

Editors: Does the evaluation interfere with mentoring or with building a trusting relationship between the participating and consulting teachers?

Melissa Joseph: When you hear that someone's going to evaluate you in your classroom and then work with you, of course it's a bit intimidating. You wonder: Is she going to see my weaknesses? Is she going to be very hard on me? But the key thing to remember is that the consulting teacher is here to help you.

When I started this job, I was intimidated by the kids. Audrey pointed out that I needed to stand firm. By the middle of the school year, I felt more confident; the kids saw that and acted accordingly. My biggest fear was that the students would intimidate me and I wouldn't be able to get through my lesson. But Audrey gave me behavior management guidelines to follow. I learned to tell students: Here's your first warning; here's your second warning. On the third warning, I send you out. Most importantly, she taught me to be consistent.

Audrey also helped me see that some behavior problems arose because students were bored. She suggested ways that I could encourage them to be more creative, such as giving them short writing prompts in which they take on different roles. This was great because it built on my strength in writing. She also helped me with strategies to keep students engaged while we are reading aloud and discussing novels.

You don't learn how to handle disruptive behavior in college. If Audrey hadn't been there to help me, I don't know how I would have gotten through the year.

Audrey Fox: The length of time we work with our interns helps to alleviate some of that initial anxiety. It fades away as soon as they realize that what the consulting teacher saw them struggle with in class—keeping students on task, for example—doesn't result in a reprimand. Consulting teachers follow observations with constructive questioning, such as asking if the teacher has tried a particular strategy. Participating teachers quickly learn that, yes, we're going to see areas that need improvement, as well as strengths that need to be reinforced. But it's always followed up with help. And they see that the person working with them is a peer. I'm able to come into someone's classroom and say, "Here's something that I've tried." That helps participating teachers buy into PAR.

Melissa Joseph: PAR is a tool, something that teachers can use to be more professional and to improve their instruction and classroom management. Of course, with any tool it all depends on whether you're willing to use it to your advantage, whether you're willing to accept the suggestions to help you achieve your goals. ❑

A Teacher Wonders

Can Grading Teachers Work?

ILLUSTRATED BY PAUL ZWOLAK

By Marc Epstein

Report Card: "a card containing a report; specifically, a card, submitted by a school, exhibiting a pupil's record to his parents or guardian."

This terse definition appeared in the classic 1934 edition of *Webster's Second International Dictionary*. Until recently, if you'd ask the proverbial man on the street for a definition, I'd venture to say that the overwhelming response would closely match *Webster's*. But all that appears to be changing since school systems throughout the country are issuing report cards to schools, and now, to their faculties. The teachers' unions, purported by critics to be omnipotent, are doing their best to participate in the debate as school reformers insist on holding teachers accountable for improving educational performance.

A recent cartoon in the *Wall Street Journal* pretty much says it all: a young student presents his failing report card to his teacher and opines, "Ah, Miss Brimsley, I ask you: Which one of us has truly failed?"

A series of new assumptions in the world of educational theory have become axiomatic. As the *Time* magazine cover story on February 25, 2008, "How To Make Great Teachers," put it, "There's no magic formula for what makes a good teacher, but there is general agreement on some of the prerequisites. One is an unshakeable belief in children's capacity to learn. 'Anyone without this has no business in the classroom,' says Margaret Gayle, an expert on gifted education at Duke University."

If every child can learn, then it follows that the reason for poor student performance must lie elsewhere. In his book *Doomed to Fail*, Paul Zoch documents the steady march of public education in this country, over the past century, toward a system of teacher-centered responsibility for learning. The latest iteration of this trend is the theory that rewarding and punishing teachers based on the extent to which their students' test scores increase will solve the riddle of public education's Gordian knot.

For example, several months ago, a political fight erupted between the United Federation of Teachers in New York City and the mayor over the validity of using students' test results to determine teacher tenure. The mayor, a strong advocate of using test results to evaluate teacher performance, said, "All of us are

Marc Epstein teaches social studies at Jamaica High School in Queens, New York. He has written articles on the New York City schools over the past five years, appearing in a variety of journals and newspapers, and was a contributor to A Consumer's Guide to High School History Textbooks*, edited by Diane Ravitch.*

judged on whether or not we do a good job. And to not judge teachers the same way, it's an insult to the teachers" (*New York Sun*, April 7, 2008). In an op-ed that appeared in the *New York Daily News* (April 8, 2008), schools Chancellor Joel Klein weighed in with his support: "Research tells us that a teacher's track record in helping students learn over a few years is a powerful indicator of whether that teacher is going to help his or her students succeed over the course of a career." At the end of the day, the New York State legislature barred the use of student test scores for making tenure decisions for a two-year period, seemingly granting another victory to the allegedly "obstructionist" teachers' union, and instead created a commission to study the issue.

The fact is, using test results to judge teacher performance is much trickier than it sounds—student test scores are influenced by all sorts of things that are beyond a teacher's control. So researchers are currently working to develop a way of isolating the teacher's impact, if indeed that is even possible. (To see how far they have come and what challenges remain, read Harvard University Professor Daniel Koretz's article that starts on page 18.)

To date, the most well-known (although not the most highly regarded) approach has been crafted by William Sanders, formerly a statistician at the University of Tennessee and now a senior research fellow with the SAS Institute Inc. It's based on the Tennessee Value-Added Assessment System, and it was developed for the Tennessee Department of Education in 1992. The Sanders model ranks teachers according to how much more, or less, growth their students have made compared with the average teacher; "effective" teachers are those whose students made above average growth (by a margin considered to be statistically significant). Sanders claims that by focusing on growth, the value-added model removes socioeconomic factors that play an important role in student achievement, such as family and home environment, and that the results can help improve teaching performance. Critics of the Sanders model argue that it is far too simple. For example, they think it does not adequately account for numerous student-background, classroom, and school factors that play a role in classroom achievement. Nonetheless, the Sanders model is just one of many. They all have their strengths and weaknesses, but none is able to fully and accurately isolate the teacher's impact on student growth.

Variations of value added have been adopted throughout the country, with New York City the nation's largest and most recent school system to sign on. The city purchased a new $80 million computer tracking system (with so many glitches that, at best, it's a work in progress) to chart the progress of its 1 million students. The desire to apply a value-added system throughout our nation's schools prompts a critical question that has largely been ignored. Will this tracking method be useful in a school system like New York City's, where all sorts of data indicate that the students are very mobile?

Many struggling urban school districts (such as Chicago and Los Angeles) have been handed over to mayors, retired generals, a former governor, a federal prosecutor, corporate lawyers, and businessmen whose only experience with the educational system is their memories of their own education. These "reformers" argue that a new paradigm that measures teacher and school performance the same way it's done in the "real world" will turn our schools around. But are their memories applicable to today?

When I've looked back at my class pictures beginning with kindergarten at P.S. 139 in Rego Park, New York, I can track the physical growth of my classmates year to year because everyone, with one or two exceptions, remained in my school. In fact, I can remember only one new addition from another country, a boy from Germany named Walter who entered my fifth-grade class.

Variations of value added have been adopted throughout the country, with New York City the most recent to sign on. But will it be useful in a school system where the students are very mobile? We lose plenty of students to other states, and the students coming in are often not only from other states, but from other countries.

Also, during my years in elementary school, not a single teacher was added to or subtracted from the faculty. Under these conditions, a value-added model might have provided us with useful data regarding student progress and teacher effectiveness. But those are not today's conditions.

Like other urban areas, New York is now a city of extraordinary mobility. Students move in, and students move out, changing schools and neighborhoods and cities. A recent study conducted by New York University's Institute for Education and Social Policy tracked the progress of about 86,000 children who entered the first grade in the fall of 1995.[1] The results are startling, even though they confirm my own observations of student turnover where I teach in Jamaica, Queens. After eight years, almost 40 percent of the students had left the New York City public schools.

Douglas Harris, a University of Wisconsin-Madison researcher who develops and studies value-added models, has noted that mobility poses a major problem for value-added models because it leads to missing data. And, although we all know that, on average, highly mobile students are not identical to their less mobile peers, these models assume that data are missing at random. As he puts it, this assumption "is especially likely to be a problem in high-poverty schools where absenteeism and mobility are high and test-taking rates are lower. It is therefore a significant question whether valid value-added estimates can be made in schools with high mobility."[2] Hopefully, officials in other cities will heed Harris's warning—those in New York City have not.

In New York City, the sophisticated new computer system tracks students' scores as they move around the district, and it can link to a statewide database as well. That reduces the missing data problem, but it certainly does not eliminate it. We lose

plenty of students to other states, and the students coming in are often not only from other states, but from other countries.

But as far as I can tell, neither this nor any other concern about the validity of value-added modeling bothers city officials at all. They are boldly piloting their own, highly suspect model that uses two years of student data and judges teacher performance by considering the growth of as few as three students. At best, this is irresponsible. But wait, it gets worse: since state achievement tests are given in the middle of each school year, the growth of all students—even those who don't switch schools at all—has to be divvied up across two teachers. This model apportions the amount of growth each teacher produced according to the number of months the teacher taught that student—a tactic that is clearly a poor substitute for an exact attribution (since it's possible that, month for month, students grew more with one teacher than the other). As a teacher, this really bothers me. I don't want the credit for another teacher's good work (or the blame for another teacher's not-so-good work).

Teacher, Mentor, Tutor, Specialist

Is Any One Educator Responsible for Student Learning?

BY LINDA VALLI, ROBERT G. CRONINGER, AND KIRK WALTERS

A fundamental premise of much of the current research on teaching is that teaching quality is central to student learning. One result of this research, though not necessarily intended, has been the call to base individual teacher evaluations on contributions to student achievement gains. Given the potentially high stakes for teachers, these proposals almost always generate heated debate. While much of the debate revolves around methodological issues in using student achievement data to evaluate teacher performance, we raise an even more fundamental question, one that has received little attention from proponents of teacher accountability policies: just who is doing the teaching?

As part of a longitudinal study of the teaching of reading and mathematics, we sought to link fourth- and fifth-grade students to the individual teacher responsible for their instruction. While we recognized that students often interact with multiple adults around subject matter, the scope, forms, and duration of these interactions surprised us. As we observed the flow of students and adults in and out of classrooms, we identified a range of more complex instructional designs quite different from the traditional "egg-crate" classroom, where one teacher works with a group of students in isolation from other adults.* In our schools, instead of students having one teacher responsible for their yearly progress in a particular subject area, many students had multiple adults engaged in their instruction, especially if the students were considered part of one or more "at-risk" groups (e.g., English language learners, low-income students, or special education students).

We started asking ourselves the question, "Who (else) is the teacher?" while engaged in a multiyear study of fourth- and fifth-grade reading and mathematics classes. Our goal was to learn more about teaching practices, as well as the allocation of school resources and educational policies, that assist or hamper the acquisition of foundational skills in these two subject areas. Although the primary purpose of the study was not to examine student assignments and alternative instructional designs, we became interested in these topics at the end of the second year of data collection because it became increasingly apparent that these designs varied among schools and among classes in schools.

The schools in the study are part of one of the largest and most diverse school systems in the nation. Over 40 percent of the students are African American or Hispanic, more than 30 percent receive free or reduced-price meals, and over 20 percent have been enrolled in English for speakers of other languages (ESOL) programs. The study design called for us to identify a group of moderate- to high-poverty schools with greater than expected achievement gains in the district, and then to follow these schools and their fourth- and fifth-grade teachers for three years.

For this study of who is doing the teaching, we drew on data collected at 18 elementary schools during the 2003-04 school year: a resource survey that asked teachers about instructional assistance; teachers' class rosters and daily logs; principal interviews about resource allocations and decision-making; and conversations with teachers about resource help and student reassignments.

As we collected data in the participating schools, we found substantial variation—some anticipated, some not—in how students and teachers were linked for instructional purposes.

Education researchers and policymakers are generally aware of some of the challenges associated with isolating teacher effects on student learning. For example, there is wide recognition that teacher absences require some sharing of instructional responsibilities among teachers. Because absences are the result of everything from attendance at individualized education program (IEP) meetings or professional development activities to personal illness or maternity leave, they may involve the sharing of instructional responsibility for a small part of a school day or a significant part of a school year. Based on the daily logs kept by teachers in the study, the average amount of time that someone other than the assigned teacher had

Linda Valli is professor in the Department of Curriculum and Instruction at the University of Maryland, College Park, where Robert G. Croninger is associate professor in the Department of Education Policy and Leadership. Kirk Walters is research analyst with the American Institutes for Research. *This article is excerpted from "Who (Else) Is the Teacher? Cautionary Notes on Teacher Accountability Systems," an article that Valli, Croninger, and Walters originally published in the August 2007 issue of the* American Journal of Education. *The full text is available online, for $10, at* ***www.journals.uchicago.edu/doi/abs/10.1086/518492***.

* Dan C. Lortie, *Schoolteacher: A Sociological Study* (Chicago: University of Chicago Press, 1975).

When all is said and done, does this value-added model have any value at all? To me, it appears not only costly but ineffective and misleading. Astronomers have the luxury of examining the light that gets to earth and is captured by radio telescopes millions of light years after a star has exploded. Educators, unlike astronomers, must have data that can be readily acted on if the data are to be of any use. These data, I suggest, have so many flaws and limitations that they should not be used to evaluate teachers. ☐

Endnotes

1. M. Weinstein, J. Pakes, C. Donis-Keller, and A. E. Schwartz, "From One to Eight: A Longitudinal Portrait of the First Grade Class of 1995-1996" (IESP Policy Brief, 2008), **http://steinhardt.nyu.edu/iesp/briefs**.

2. D. Harris, "Would Accountability Based on Teacher Value-Added Be Smart Policy?" (paper for the National Conference on Value-Added Modeling, April 22-24, 2008).

responsibility for instruction in reading and mathematics due to absences was roughly 7 percent.

But even when teachers are present, other factors confound a clear linkage between student achievement and teacher performance. Student mobility is one factor. With the average mobility rate in these schools at 20 percent, a significant number of students in the study would have had a teacher from another school responsible for part of their instruction during the course of the school year. An additional complication arises from the public notification requirements introduced by No Child Left Behind (NCLB). In the district we studied, the testing schedule for the purposes of NCLB ran from March to March so that parents could be provided with test results prior to the beginning of the next school year. This meant that every teacher in the study shared responsibility for achievement gains with at least one teacher from the previous year. Given the district's 9.5-month school calendar, this amounts to roughly one-quarter of students' "tested" instructional time.

Even when students stayed in the classroom, someone other than the classroom teacher could have had responsibility for their instruction. We observed classrooms where the teacher of record consistently worked with one reading group while instructional assistants worked with others, where student teachers took over a substantial proportion of instructional responsibilities, and where a staff developer took over part of the lesson to demonstrate a teaching strategy.

There were also numerous instances where students were assigned to a specific reading or mathematics class for part of the period and sent to an ESOL or resource teacher for the rest of the period, or where a student spent the entire instructional period with the classroom teacher and received an additional reading or mathematics lesson during another part of the day with a different teacher. Homeroom teachers, who were not the reading or mathematics teacher of record, gave students work during the homeroom period targeting skills or concepts presumed to be on the annual state assessment, and computer teachers pulled small groups of students from the classroom to work on writing assignments in the computer lab. In one school, literacy instruction was divided into two separate classes, with one teacher instructing students in reading and a different teacher instructing them in writing.

In addition, we observed a surprising amount of fluidity in teacher-student assignments in some of the schools. Although the principal generally made the formal assignments at the beginning of the school year, grade-level teams sometimes adjusted these assignments, with or without the principal's knowledge. For example, grade-level teachers might pair up and switch students for a particular instructional unit and then switch students back again. In one mathematics class, two teachers were originally assigned to co-teach a large group of students, but later in the year the group was split into two separate classes.

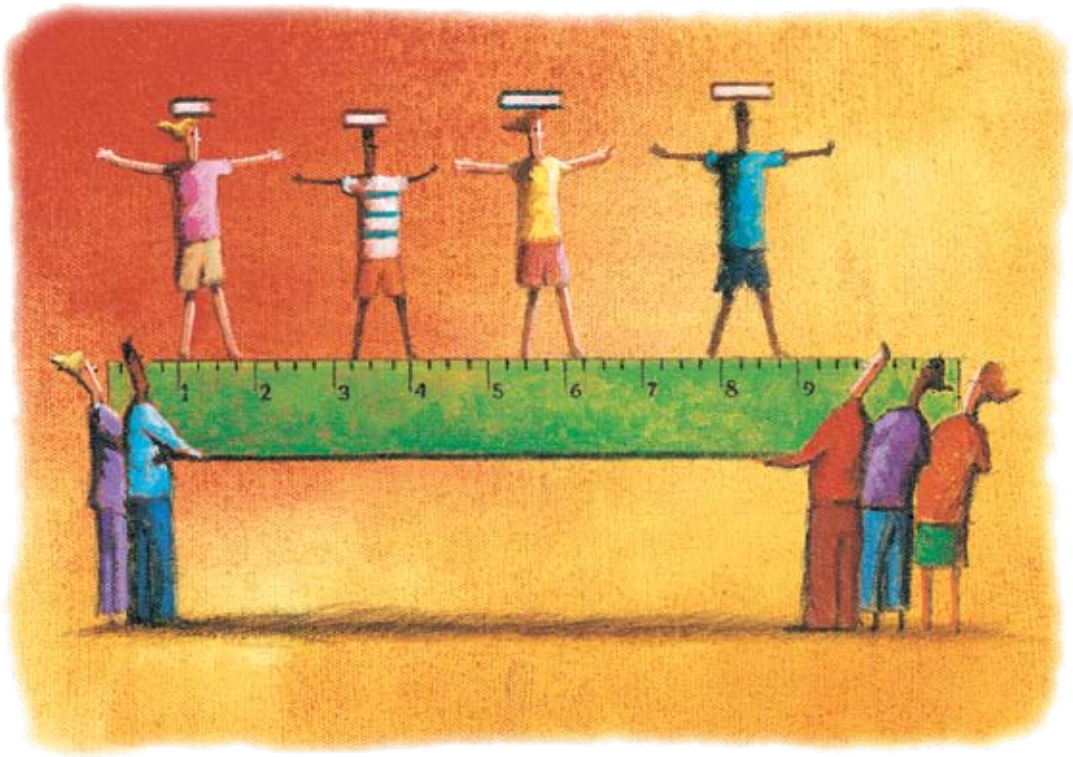
* * *

These findings raise questions about both the feasibility and desirability of teacher accountability systems based on student achievement data. In this era of high-stakes accountability, caution must be taken to ensure that responsibility for student learning is accurately attributed. Our analysis of these 18 schools, 69 teachers, and over 1,500 students suggests that less responsibility rests with the formally assigned classroom teacher than we initially assumed or that past studies led us to anticipate. It makes little sense to have an individual accountability model when multiple actors have a role in student learning.

Furthermore, our understanding of the potential benefits of other reform efforts tempers whatever enthusiasm we might have had for the teacher accountability movement. Even if more sophisticated statistical methods eventually make possible a more accurate attribution of teaching impact for multiple actors, this may not be a desirable direction for educational policy. It can too easily derail other efforts to support high-quality teaching and learning, including the promotion of professional learning communities and the flexible, coordinated use of trained teacher resources. This does not mean that efforts to understand and improve teaching quality are ill-conceived, only that, in many instances, teaching is a collective rather than solely individual pursuit. Education policies and teacher accountability systems need to reflect this reality.

A Measured Approach

Value-Added Models Are a Promising Improvement, but No One Measure Can Evaluate Teacher Performance



ILLUSTRATED BY PAUL ZWOLAK

By Daniel Koretz

Suppose you and I teach fifth grade—as I did many years ago—but we teach in very different settings. Our students are different: perhaps yours enter fifth grade with lower levels of achievement, or you have more students with limited proficiency in English. Their previous teachers were not similar: perhaps those who taught my students were more skilled. On average, my students have more highly educated parents than yours. Our schools have different levels of resources, and the peer culture and community support for education are dissimilar. But our students do have one thing in common: at the end of the school year, our students will take the same achievement tests, and policymakers would like to use their scores to judge how effective we both were. How fairly can that be done, given our very different situations?

The education policy community is abuzz with interest in value-added modeling as a way to estimate the effectiveness of schools and especially teachers—even those with very different students, in very different settings. Value-added approaches are widely believed to be superior to the common alternatives as a way of estimating the performance of schools and teachers. But just how well do value-added models serve this role? There is no doubt that value-added models are superior in some important ways, but they are no silver bullet. Value-added models provide important information, but that information is error-prone and has a number of other important limitations. Moreover, these methods are still under development, and the various approaches now in use do not always paint the same picture. Value-added estimates can be an important part of an evaluation of teachers and schools, but they are not sufficient by themselves for this purpose.

Although there has been intense discussion of the strengths and limitations of the value-added approach among research-

Daniel Koretz is professor of education at Harvard University. He founded and chairs the International Project for the Study of Educational Accountability and is a member of the National Academy of Education. His research focuses on the effects of high-stakes testing, including effects on schooling and the validity of score gains, and the design and evaluation of test-focused educational accountability systems. Previously, he taught emotionally disturbed students in public elementary and junior high schools. *He wishes to thank Daniel McCaffrey, J.R. Lockwood, and Laura Hamilton—colleagues with whom he worked on RAND's evaluation of value-added modeling—for their helpful comments on an earlier draft.*

ers, too little discussion has taken place in the education policy community. This may stem from the tremendous technical complexity of most value-added approaches, which render them seemingly incomprehensible to most people, or from policymakers' hope for a relatively simple way of evaluating teachers and schools, or both. Yet without this discussion, we are not likely to use value-added modeling in an appropriate and productive way. This article describes some of the key issues raised by value-added modeling and concludes with some suggestions for its use. Many of the issues are similar regardless of whether schools or teachers are evaluated, and I touch on both, but I focus especially on the evaluation of teachers.

How Value Added Improves on the Status Quo

Most test-based accountability programs in the United States have used one of three approaches for evaluating student achievement. *Status models* are based simply on the scores of a group at one time. For example, the average performance of a school's fourth graders, or the proportion of fourth graders who exceed a standard such as "proficient," can be compared with an expected level or with the results from other schools. *Cohort-to-cohort change models* are based on the change in statistics such as these over time. For example, the percentage of fourth graders considered proficient this year can be contrasted with the comparable statistic from last year to see which schools have attained an expected degree of improvement. The federal education law, No Child Left Behind (NCLB), is a hybrid of these two approaches. For most schools, NCLB functions as a status model: in any given year, the performance of the school is compared with the state's annual measurable objective for that year. However, the objective increases every year (on its way to the goal of 100 percent proficient by 2014), which creates pressures similar to that found in a cohort-to-cohort change system. In addition, NCLB's safe harbor provision is a true cohort-to-cohort change approach.

In contrast to both of these, *value-added models* (VAMs) are based on the growth individual students achieve during a year of schooling. If I were still a fifth-grade teacher, a status model would evaluate me based on my students' performance at the end of this year, and a cohort-to-cohort change model would judge me based on the difference between the end-of-year scores of my fifth graders this year and those I had the year before. Under a VAM, I would be rated on the basis of my students' gains during their year with me; I would be evaluated favorably if they showed more growth than whatever comparison policymakers decided to use (which might be the average of other teachers in my district or state, or some pre-established amount), even if my students' performance when entering my class was so weak that their scores at the end of fifth grade remained low.

Unfortunately, the term "value added" is used to represent two very different quantities. The first is students' total growth—how much their achievement increased, *for whatever reason*, during their fifth-grade year with me. The second is how much *my efforts* contributed to that growth—how much "value" I added. Because many factors other than teachers' work contribute to (or impede) growth, these two quantities can be quite different. I'll use the term *value added* to refer to both for now, but I'll return to this distinction later.

In test-based accountability systems, value-added approaches offer three very important advantages compared with status models and cohort-to-cohort change models. First, at least in theory, VAMs measure the right thing, which neither status nor cohort-to-cohort change models do. A sensible accountability system, for teachers or for any other professionals, holds people accountable for what they can control. Teachers should be held accountable for what they contribute to their students' growth, not for the accumulated knowledge and skills (or lack thereof) that students bring with them to the first day of class.

The education policy community is abuzz with interest in value-added modeling as a way to estimate the effectiveness of schools and especially teachers. Value-added models provide useful information, but that information is error-prone and has a number of other important limitations.

While adjusting for students' achievement levels when they enter the grade would be a clear and important improvement over cohort-to-cohort change and status models, it is not enough to get us a true estimate of "value added." The ideal is to adjust not only for students' prior achievement levels, but rather for their *expected growth trajectories*. To better understand this, let's go back to the example of fifth grade, and let's add the condition that you and I are equally effective teachers. This time, let's assume that, for whatever reason, you are given a class of high achievers, with very few students reading below grade level and many reading several years above grade level. In contrast, I draw—as I did in actuality, many years ago—a class with many very poor readers, some several full years below grade level. (So far below, in fact, that many still struggled with decoding and read letter by letter.) Would these two groups gain reading skills at the same rate if they had equally effective teachers? Should I be judged less effective than you if my students gained less in reading skills during the fifth grade than yours? Most experienced teachers, I suspect, would say no. To make the comparison truly fair, one would want the system to adjust for *differences in the growth* that these two very dissimilar groups would show during fifth grade if they were given equally high-quality schooling.

The achievement level of students when they enter a grade

reflects the cumulative effects of many factors, both educational and not. Some of these factors will persist after the students enter your class and will tend to push them toward a growth trajectory similar to that which they showed before. Some of these are characteristics of the students themselves, such as disabilities, health conditions, and simple differences in aptitude. Some are characteristics of their families or communities. For example, my own children attended school in a neighborhood in which many parents either hired tutors or retaught material themselves if their children encountered difficulties (as I did when my son encountered difficulties with his mathematics homework)—which increased their children's rate of growth and gained the schools some credit they did not actually deserve. The combined effects of these influences make some students much easier to teach than others. I have taught in settings ranging from special education elementary school classes to doctoral-level university courses, and this variation in students has been striking in every class I have taught.

Therefore, some current VAMs try to adjust for differences in students' expected growth trajectories by taking into account several years of prior achievement, not just scores from the year before entry to a class. By evaluating several years of scores, the models indirectly take into account persistent noneducational factors that influence students' rate of growth, and some approaches also incorporate some of these factors directly into the model.

This brings us to the second main advantage of VAMs: they can do a substantially better job than status models or cohort-to-cohort change models of controlling for differences among students that would otherwise be confounded with the effects of teaching. Currently, there is a great deal of argument among experts about how well VAMs do this—how close they come to estimating the value added by teachers rather than just estimating student growth. For reasons that I will explain below, we cannot be confident that value-added models pare away all of the growth attributable to other factors in order to reveal the pure effects of teaching. Nonetheless, in general, VAMs do a better job of adjusting for other influences on achievement than do the typical status or cohort-to-cohort approaches.

The final major advantage of VAMs is that they reveal substantial differences among classrooms and schools in students' performance. We all have known superb teachers and teachers who are struggling, so it is reasonable to expect a measure of student performance to show substantial variations. Test scores show great variation among schools, but research has often found that after adjusting for factors such as background characteristics, relatively little variation—implausibly little, some observers would say—remains. In contrast, VAM estimates often show the sizeable differences among teachers and schools that many would expect.*

Difficulties in Using Value-Added Models for Accountability

Applying VAMs to the evaluation of schools and teachers is not straightforward, and some of the issues debated by experts, while important, seem simply impenetrable to most people other than statisticians and psychometricians. This in itself is a drawback, as it's certainly preferable for educators, parents, policymakers, and the like to understand how their teachers and schools are being evaluated. Fortunately, many of the most important complications can be reduced to the following six simple questions, each of which I'll briefly discuss: (1) What are we measuring?

Value-added models can do a better job than the alternatives of controlling for differences among students that would otherwise be confounded with the effects of teaching. But we cannot be confident that value-added models pare away all of the growth attributable to other factors in order to reveal the pure effects of teaching.

(2) How do we measure it? (3) How precise can we be? (4) How certain are we about how to model gains? (5) How well do we adjust for other influences on achievement growth? (6) How does score inflation affect value-added models?

1. What are we measuring?

It is essential to keep in mind a warning offered by some of the progenitors of achievement testing more than half a century ago: standardized achievement tests can only measure a subset of the critically important goals of education. First, they measure only achievement, not motivation, curiosity, creativity, and the ability to work well in groups. Second, most testing systems measure achievement in only a subset of the subject areas with which we should be concerned. Third, within the tested subject areas, they measure only a subset of the important knowledge and skills. Some important outcomes are very difficult or impractical to test with standardized, externally imposed tests. The information yielded by standardized tests can be tremendously valuable, but it is nonetheless seriously incomplete, and therefore scores taken alone cannot provide a comprehensive evaluation of perfor-

* In the current context of NCLB, another advantage is that most value-added models take into account every student's progress. In contrast, NCLB and most state accountability systems focus primarily on the percentages of students reaching a proficient standard, which renders progress by most students—those well below or well above the standard—invisible and unimportant. As I explain in my new book *Measuring Up* (see chapter 8), this is only one of many serious drawbacks of reporting student achievement only in terms of performance standards. However, this advantage is not inherent to VAMs. There is no reason why cohort-to-cohort change or status models need to focus on the percentages of students reaching a standard rather than on the performance of all students.

mance. (To better understand this concern, see the sidebar from *Measuring Up* on page 22.)

Far from circumventing this problem, value-added models may exacerbate it. The VAMs we use today require that growth in achievement be cumulative across grades. We want to know how far a student has progressed in learning mathematics by the end of grade 4, so that we can evaluate how much her knowledge has increased by the end of grade 5. This requires *vertically scaled* tests: tests that place performance in adjacent grades on a single scale.† The more dissimilar the content of instruction is from grade to grade, the less plausible this approach is. Vertically scaled tests are commonplace in reading comprehension and certain areas of mathematics, but they may not be practical in science or social studies, even in the elementary and middle grades. More subtle, but also important, is that using VAMs may constrain what we test within a subject as well. The more grade-specific the important content in one subject is, the less practical it becomes to build defensible vertically scaled tests. Therefore, reliance on VAMs may encourage focusing on a subset of important subjects and narrowing the focus within subjects to the material most amenable to vertical scaling.

2. How do we measure achievement?

Although many people believe that tests are direct and simple measures of achievement, they are anything but. A test is only a small sample from a large "domain" of knowledge and skill, and performance on the tested sample—the test score—is only valuable to the extent that it provides a good estimate of mastery of the entire domain. (These issues are explored in the sidebar on page 22.) Constructing a test entails a long series of decisions, both substantive and technical. Some of these decisions, such as the choice of a mathematical model for creating a scale, are arcane, but they matter: they can substantially affect the estimates of gains that are provided by value-added models. I'll give three examples.

The first is the selection of content. Consider middle school mathematics. In many middle schools, there is considerable tracking in mathematics, and there are likewise curricular differences among schools. Some seventh graders are studying algebra, while others are still focused on arithmetic. Suppose you and I are equally effective seventh-grade math teachers. You are teaching a class in which a good deal of time is devoted to algebra, while I am teaching one focused primarily on arithmetic. Suppose also that our state uses a test that focuses on basic skills. What will value-added models say about us? You lose: much of the progress you make with your students will not be captured by the test because it does not include algebra. The technical term you may see for this is *dimensionality*. Most tests measure multiple aspects or dimensions of performance, although they provide a summary score combining all of them. The closer the mix of tested dimensions is to your curriculum, the more effective you will seem.

The second testing issue is scaling: deciding on a set of numbers to represent performance. Most value-added approaches assume an *interval scale*, such that any given increment, say, 20 points, means the same improvement in achievement at any level of the scale (so, for example, an increase from 120 to 140 represents the same amount of growth as an increase from 200 to 220). Most people don't give this concern much thought, since most of the measures we use in daily life, such as pounds, feet, and temperature, are interval scales. Unfortunately, test scores do not have this handy property: we would like an interval scale, but most of the time we don't know whether we have one. We can't be confident that, for example, an increase from 500 to 540 on the SAT mathematics test represents the same amount of gain as an increase from 700 to 740. Worse, different scales do not necessarily agree in this regard. A high-achieving student and a low-achieving student who appear to have gained the same amount on one scale may show different amounts of growth on another.

For many practical purposes, this uncertainty does not matter much. For example, it has been shown that many of the commonly used scales correspond reasonably well in this respect, provided that the comparison is restricted to one grade and year, and to students who are not dramatically different in performance. However, it clearly can matter with VAMs. For example, some scales will show the performance of high achievers and low achievers diverging as they progress through the grades, while others show the reverse, and yet others show the two groups keeping pace with each other. This creates a distressing uncertainty in the results of value-added models when the groups compared start out at substantially different levels of achievement. (I'll return to this at the end, when I offer some suggestions about using VAMs sensibly.)

The final example is the timing of testing. Most states test once a year, near (but not at) the end of the school year. Therefore, the growth attributed to a teacher excludes the final weeks or months of the school year and includes both the final period in the previous year (with the previous teacher) and summer vacation. Particularly given evidence that students show different patterns of growth or loss during the summer, these problems of timing are worrisome.‡ Although there are some statistical simulations suggesting that the effects of this less-than-optimal timing are usually not great, the jury is still out, and there may be some circumstances in which this is an appreciable source of bias in the ranking of teachers or schools.

3. How precise can we be?

Years ago, fresh out of graduate school, I wrote testimony for a congressional committee in which I referred to the "margin of error" in my estimate of the impact of a program the committee was considering terminating. This angered the chair of the committee, who glowered at the person giving the testimony—unfortunately, my boss—and said, "What is this 'margin of error' stuff? Doesn't it mean that you don't know what the hell you're talking about?" Well, in a sense, yes, although he was overstating the problem. While the chair wanted certainty, no one could honestly give it to him: all statistical estimates are subject to some uncertainty or imprecision, and this includes test scores and the results of models that use them. Terms such as "margin of error" or the more specific "standard error" are just our tools for quantifying how much imprecision remains.

To start, we have to distinguish between *error* and *bias*. In

† A few value-added models loosen this requirement slightly, but these exceptions do not contradict the points made here. There are also statistical approaches for estimating the value added by individual teachers that are not based on prior growth in the same subject area, but we are not considering those here.

‡ For more on summer learning loss, see "Keep the Faucet Flowing" in the Fall 2001 issue of *American Educator*, online at **www.aft.org/pubs-reports/american_educator/fall2001/faucet.html**.

educational testing, as in most of quantitative science, "error" has a narrower meaning than it does in common parlance. If you buy a cheap bathroom scale, it may simply be inconsistent, so that your weight seems to be different each time you step on it, but not systematically too high or too low. This inconsistency is error. On the other hand, your bathroom scale could be systematically wrong, so that it consistently tells you that you are lighter than you really are. In educational testing, this systematic inaccuracy is called *bias*, not error. If a student's score is consistently too low, as may happen in the case of students not fully proficient in English, that would constitute bias; but if a student's score is sometimes too low and sometimes too high, that would be error.

Even if they are entirely unbiased, estimates based on test scores inevitably entail error. In fact, both bias and error are concerns when value-added models are used to evaluate teachers or schools. I'll discuss error here and return to bias a bit later.

Error is of two analogous types that have different sources: *sampling error*, which is more familiar to most people, and *measurement error*.* *Sampling error* stems from the selection of particular individuals from whom data will be collected. In the case of educational accountability, sampling error arises because a teacher is given a different sample of students every year, and, as one teacher put it in a study years ago, "there are good crops and bad crops." Your scores—and your apparent "effectiveness"—will fluctuate as a result of these differences in samples. These fluctuations are particularly pronounced for small groups

* I provide a more thorough explanation of bias, measurement error, and sampling error in *Measuring Up*.

Measuring Up

What Educational Testing Really Tells Us

Educational testing is ubiquitous in America, and its importance is hard to overstate. Tests have a powerful influence on public debate about many social concerns, such as economic competitiveness, immigration, and racial and ethnic inequalities. And achievement testing seems reassuringly straightforward and commonsensical: we give students tasks to perform, see how they do on them, and thereby judge how successful they or their schools are.

This apparent simplicity, however, is misleading.

Test scores do not provide a direct and complete measure of educational achievement. Rather, they are incomplete measures, proxies for the more comprehensive measures that we would ideally use, but that are almost always unavailable to us. There are two reasons for the incompleteness of achievement tests. The first, which has been stressed by careful developers of standardized tests for more than half a century, is that these tests can measure only a subset of the goals of education. Some goals, such as the motivation to learn, the inclination to apply school learning to real situations, the ability to work in groups, and some kinds of complex problem solving, are not very amenable to large-scale standardized testing. Others can be tested, but are not considered a high enough priority to invest the time and resources required. The second reason for the incompleteness of achievement tests—and the one that I will focus on here—is that even in assessing the goals that we decide to measure and that can be measured well, tests are generally very small samples of behavior that we use to make estimates of students' mastery of very large domains of knowledge and skill.

The accuracy of these estimates depends on several factors, one of the most important being careful sampling of content and skills. For example, if we want to measure the mathematics proficiency of eighth graders, we need to specify what knowledge and skills we mean by "eighth-grade mathematics." We might decide that this subsumes skills in arithmetic, measurement, plane geometry, basic algebra, and data analysis and statistics, but then we would have to decide which *aspects* of algebra and plane geometry matter and how much weight should be given to each component (e.g., do students need to know the quadratic formula?). Eventually, we end up with a detailed map of what the test should include, often called "test specifications" or a "test blueprint," and the developer writes test items that sample from it.

But that is just the beginning. The accuracy of a test score depends on a host of often arcane details about the wording of items, the wording of "distractors" (wrong answers to multiple-choice items), the difficulty of the items, the rubric (criteria and rules) used to score students' work, and so on. The accuracy of a test score also depends on the attitudes of the test takers—for example, their motivation to perform well. It also depends, as we shall see later, on how schools prepare students for the test. If there are prob-

because there is less opportunity for the characteristics of individual students to cancel each other out. Thus, the smaller the group, the greater the sampling error, and the greater the uncertainty in the group's test scores—or in the estimates of value added based on them.

Measurement error is different: it affects even the score of a single student and reflects inconsistencies from one instance of measurement to another. Students who take the SAT multiple times, for example, generally see a fluctuation in their scores from one time to the next because of measurement error. As explained in the sidebar (below) from *Measuring Up*, there are three primary sources of measurement error: the selection of specific test items in constructing the test, fluctuations in the student's performance from day to day, and inconsistencies in scoring.† Some states and districts now take measurement error into account when reporting scores, telling parents that the best estimate of a student's performance falls within a range surrounding her obtained score.

The score reports used in accountability systems are subject to both measurement error and sampling error. As a result, one can't take the precise score obtained for a school or classroom at face value. Rather, the score is an estimate, and the true value lies within a band of uncertainty that surrounds the estimate obtained. (This is no different from the polls you see in the news-

(Continued on page 26)

† *Reliability* is a function of error: a perfectly reliable score would be error-free (in most cases, an impossibility), while a completely unreliable score would represent nothing but error.

lems with any of these aspects of testing, the results will provide misleading estimates of students' mastery of the larger domain.

A failure to grasp this fact is at the root of widespread misunderstandings—and misuses—of test scores. It has often led policymakers astray in their efforts to design productive testing and accountability systems. By placing too much emphasis on test scores, they have encouraged schools to focus instruction on the small sample actually tested rather than the broader set of skills the mastery of which the test is supposed to signal.

To make the principles of testing concrete, let's construct a hypothetical test. Suppose that you publish a magazine and have decided to hire a few college students as interns to help out. You receive a large number of applicants and have decided that one basis for selecting from among them is the strength of their vocabularies. How do you determine that? Conversations with them will help, but may not be sufficient because they are not uniform: a conversation with one applicant may afford more opportunities for using advanced vocabulary than a conversation with a second one. So you decide to construct a standardized test of vocabulary.* You would then confront a serious difficulty: although many teachers and parents may find this fact remarkable in the light of their own experience, the typical adolescent has a huge working vocabulary. Clearly, you will have to select a sample of words to put into your test. In practice, you can get a reasonably good estimate of the relative strengths of applicants' vocabularies by testing them on a small sample of words, if those words are chosen carefully. Assume you will use 40 words, which would not be an unusual number in an actual vocabulary test.

The box below gives the first few words from three lists that you could use to select words for your test.

A	B	C
siliculose	bath	feckless
vilipend	travel	disparage
epimysium	carpet	minuscule

Which list would you use? Clearly not list A, which comprises specialized, very rarely used words. Everyone would receive a score of zero or nearly zero, and that would make the test useless: you would gain no useful information about the relative strengths of their vocabularies. List B is no better. Everyone would obtain a perfect or nearly perfect score. Therefore you would construct your test from list C, which comprises words that some applicants would know and others not.

In this example, the fact that a test is merely a sample of a larger domain is clear. But is sampling always as serious a problem as it is in this contrived example? For the most part, yes.† The tests that are of interest to policymakers, the press, and the public at large entail substantial sampling because they are designed to measure sizable domains, ranging from knowledge acquired over a year of study in a subject to cumulative mastery of material studied over several years.

Returning to the vocabulary test: what would have happened if you had chosen words differently, while keeping them at the same level of difficulty? To make this concrete, assume that you selected all three of the words shown in list C, and that I was also constructing a vocabulary test, but I dropped *feckless* and used *parsimonious* instead. For the sake of discussion, assume that these two words are equally difficult.

What would be the impact of administering my test rather than yours? Over a large enough number of applicants, the average score would not be affected at all, because the two words in question are equally difficult. However, the scores of some individual students would be affected. Even among students with comparable vocabularies, some would know *feckless* but not *parsimonious*, and vice versa.

This illustrates one source of *measurement error*, which refers to inconsistency in scores from one measurement to the next. To some degree, the ranking of your applicants will depend on which words you select from list C, and if you tested applicants repeatedly using different versions of your test, the rankings would vary a little. Another source of measurement error is the fluctuation over time that would occur even if the items were the same. Students have good and bad days. For example, a student might sleep well before one test date but be too anxious to sleep well another time. Or the examination room may be overheated one time but not the next. Yet another source of measurement error is inconsis-

* People incorrectly use the term *standardized test*—often with opprobrium—to mean all sorts of things: multiple-choice tests, tests designed by commercial firms, and so on. In fact, it means only that the test is uniform: that is, that all examinees face the same tasks, administered in the same manner, and scored in the same way. The motivation for standardization is to avoid irrelevant factors that might distort comparisons among individuals.

† There are tests that are not samples of a larger domain. For example, a teacher may want to know whether her class has mastered the list of vocabulary words presented in the past week. She would not be trying to draw any conclusions about students' overall vocabularies, and she would be happy indeed if most students got most of the words right.

tencies in the scoring of students' responses.

Obviously, it's important to try to keep measurement error to a minimum—and that's why test developers are so concerned with *reliability*. Reliable scores show little inconsistency from one measurement to the next—that is, they contain relatively little measurement error. Reliability is often incorrectly used to mean "accurate" or "valid," but it properly refers only to the consistency of measurement. A measure, including a test, can be reliable but inaccurate—such as a scale that consistently reads too high.

So when all is said and done, how justified would you be in drawing conclusions about vocabulary from the small sample of words on your test? This is the question of *validity*, which is the single most important criterion for evaluating achievement testing. In public debate, and sometimes in statutes and regulations as well, we find reference to "valid tests," but tests themselves are not valid or invalid. Rather, inferences based on test scores are valid or not. A given test might provide good support for one inference, but weak support for another. For example, a well-designed end-of-course exam in statistics might provide good support for inferences about students' mastery of basic statistics, but very weak support for conclusions about mastery of mathematics more broadly. The question to ask is: how *well supported* is the conclusion?

None of the preceding is particularly controversial. These fundamentals of testing may not be well known outside the testing community, but inside that community they are widely agreed upon. The next and final step in this hypothetical exercise, however, is contentious indeed.

Suppose you are kind enough to share with me your test of 40 words. And suppose I intercept every single applicant en route to taking your test, and I give each one a short lesson on the meaning of every word on your test. What would happen to the validity of inferences you might want to base on your test scores?

Clearly, your conclusions about which applicants have stronger vocabularies would now be wrong. Most students would get high scores, regardless of their actual vocabularies. Students who paid attention during my mini-lesson would outscore those who did not, even if their actual vocabularies were weaker. Mastery of the small sample of 40 words would no longer represent variations in the students' actual working vocabularies.

This last step—teaching the specific content of the test, or material close enough to it to undermine the representativeness of the test—illustrates the contentious issue of *score inflation*, which refers to increases in scores that do not signal a commensurate increase in proficiency in the domain of interest. Inflation of scores in this case did not require any flaw in the test, and it did not require that the test focus on unimportant material. The 40 words were fine. My response to those 40 words—my form of test preparation—was not.

In real-world testing programs, issues of score inflation and test preparation are far more complex than this example suggests. So let's set aside our vocabulary test and take a closer look at what I believe should be a very serious concern among educators and policymakers: how to prepare for tests.

Test preparation has been the focus of intense argument for many years, and all sorts of different terms (like "teaching the test" and "teaching to the test") have been used to describe both good and bad forms. I think it's best to ignore all of this and to distinguish instead between seven different types of test preparation: (1) working more effectively, (2) teaching more, (3) working harder, (4) reallocation, (5) alignment, (6) coaching students, and (7) cheating.

The first three are what some proponents of high-stakes testing want to see. Clearly, if educators find ways to work more effectively—for example, developing better curricula or teaching methods—students are likely to learn more. Up to a point, if teachers spend more time teaching, achievement is likely to rise. The same is true of working harder in school, although this can be carried too far. For example, it is not clear that depriving young children of recess, which some schools are now doing in an effort to raise scores, is effective, and in my opinion it is undesirable regardless. Similarly, if students' workload becomes excessive, it may interfere with learning and may also generate an aversion to learning. But if not carried to excess, these three forms of test preparation can be expected to produce real gains in achievement that would appear not only in the test scores used for accountability, but on other tests and outside of school as well.

At the other extreme, cheating is unambiguously bad. But what about reallocation, alignment, and coaching? All three can produce real gains, score inflation, or both. Reallocation refers to shifting instructional resources—classroom time, homework, parental nagging, whatever—to better match the content of a specific test. A quarter century of studies confirm that many teachers reallocate instruction in response to tests. And some studies have found that school administrators reassign teachers to place the most effective ones in the grades in which important tests are given.[1]

Is reallocation good or bad? Does it generate real gains in achievement or score inflation? This depends on what gets more emphasis, *and what gets less*. Some reallocation is desirable and is one of the goals of testing programs. For example, if a ninth-grade math test shows that students do relatively poorly in solving basic algebraic equations, one would want their teachers to put more emphasis on such equations. The rub is that devoting more resources to topic A entails fewer resources for topic B.

Scores become inflated when topic B—the material that gets less emphasis as a result of reallocation—is also an important part of the domain. If teachers respond to a test by de-emphasizing

material that is important to the domain but is not given much weight on the particular test, scores will become inflated. Performance will be weaker when students take another test that places emphasis on those parts of the domain that have been neglected.

Alignment is a lynchpin of policy in this era of standards-based testing. Tests should be aligned with standards, and instruction should be aligned with both. And alignment is seen by many as insurance against score inflation, but this is incorrect. Alignment is just reallocation by another name. Whether alignment inflates scores also depends on the importance of the material that is de-emphasized. And research has shown that standards-based tests are not immune to this problem. These tests are still limited samples from larger domains, and therefore focusing too narrowly on the content of the specific test can inflate scores.

Coaching students refers to focusing instruction on small details of the test, many of which have no substantive meaning. Coaching need not inflate scores. If the format or content of a test is sufficiently unfamiliar, a modest amount of coaching may even increase the validity of scores. For example, the first time young students are given a test that requires filling in bubbles on an answer sheet that is going to be scored by a machine, it is worth spending a very short time familiarizing them with this procedure before they start the test.

Most often, however, coaching students either wastes time or inflates scores. A good example is training students to use a process of elimination in answering multiple-choice questions. A *Princeton Review* test-prep manual urges students to do this because "it's often easier to identify the wrong answers than to find the *correct* one."[2] What's wrong with this? The performance gains generated depend entirely on using multiple-choice items. Of course, when students need to apply their knowledge in the real world outside of school, the tasks are unlikely to appear in the form of a multiple-choice item.

This example shows that inflation from coaching is in one respect unlike inflation from reallocation. Reallocation inflates scores by making performance on the test unrepresentative of the larger domain, but it does not distort performance on the material tested. (If I taught applicants the vocabulary words on your test, they would know those words—but their scores on the test would not be good estimates of their overall vocabulary knowledge.) In contrast, coaching can exaggerate performance on the tested material. In the example just given, students who are taught to use the process of elimination as a method for "solving" certain types of equations will know less about those types of equations than their performance on the test indicates.

So what distinguishes good and bad test prep? The acid test is whether the gains in scores produced by test preparation truly represent meaningful gains in student achievement. We should not care very much about a score on a particular test. What we should be concerned about is the knowledge and skills that the test score is intended to represent. Gains that are specific to a particular test and that do not generalize to other measures of the domain and to performance in the real world are worthless.

* * *

This brings me to a final, and politically unpalatable, piece of advice: we need to be more realistic about using tests as a part of educational accountability systems. Systems that simply pressure teachers to raise scores on one test (or one set of tests in a few subjects) are not likely to work as advertised, particularly if the increases demanded are large and inexorable. They are likely instead to produce substantial inflation of scores and a variety of undesirable changes in instruction, such as excessive focus on old tests, inappropriate narrowing of instruction, and a reliance on test-taking tricks.

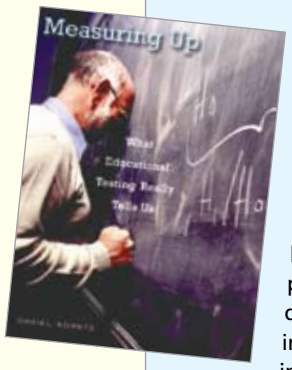
I strongly support the goal of improved accountability in public education. I saw the need for it when I was an elementary school and junior high teacher, many years ago. I saw it as the parent of two children in school. Nothing in more than a quarter century of education research has led me to change my mind on this point. And it seems clear that student achievement must be one of the most important things for which educators and school systems should be accountable. However, we need an effective system of accountability, one that maximizes real gains and minimizes bogus gains and other negative side effects. Even a very good achievement test will leave many aspects of school quality unmeasured. Some hard-core advocates of high-stakes testing disparage this argument as "anti-testing," but it is a simple statement of fact, one that has been recognized within the testing profession for generations.

So how should you use scores to help you evaluate a school? Start by reminding yourself that scores describe some of what students can do, but they don't describe all they can do, and they don't explain why they can or cannot do it. Use scores as a starting point, and look for other evidence of school quality—ideally not just other aspects of student achievement but also the quality of instruction and other activities within the school. And go look for yourself. If students score well on math tests but appear bored to tears in math class, take their high scores with a grain of salt, because an aversion to mathematics will cost them later in life, even if their eighth-grade scores are good.

Sensible and productive uses of tests and test scores rest on a single principle: don't treat "her score on the test" as a synonym for "what she has learned." A test score is just one indicator of what a student has learned—an exceptionally useful one in many ways, but nonetheless one that is unavoidably incomplete and somewhat error-prone.

–D.K.

This sidebar was adapted from Daniel Koretz's new book, *Measuring Up: What Educational Testing Really Tells Us*. Detailed but nontechnical, the book addresses the common misunderstandings and misuses of standardized tests, and offers sound advice for using tests responsibly. To learn more, go to **www.hup.harvard.edu/catalog/KORMAK.html**. *Measuring Up*,



Endnotes

1. For a good overview of some of the most important research on teachers' and principals' responses to testing, see Brian M. Stecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education*, ed. Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, CA: Rand, 2002), **http://www.rand.org/pubs/monograph_reports/MR1554**.

2. Jeff Rubenstein, *Princeton Review: Cracking the MCAS Grade 10 Math* (New York: Random House, 2000), 15.

(Continued from page 23)

paper: they are usually reported with a "margin of error" of plus or minus a few percentage points, which is their band of uncertainty.) This inevitable error is one of several reasons why no single measure should be used to make an important decision. Even if a measure is entirely unbiased, any single test score may be too high or too low, sometimes by a considerable amount.

Error affects all accountability approaches—status, cohort-to-cohort, and value-added models. There is still disagreement among experts about the precise amount of error in different VAMs, but there is no doubt that it is a serious problem indeed, particularly when the model is applied to individual teachers (since they have a limited number of students, the sample size is small, and sampling error is large). To rank teachers based on VAMs, we would need very small errors, and research to date suggests that we cannot yet reach that threshold. We may be able to identify some teachers whose students show higher- or lower-than-average gains, but it does not seem that we can be much more precise than that. For example, if one wanted to rebuke or intervene with teachers in the bottom decile in terms of growth or reward those in the top decile, we would often select the wrong teachers.

There are two ways to lessen this problem (although there is no way to eliminate it entirely). One is to add more data, which one might do by combining each teacher's or school's results from several years (e.g., instead of just looking at my value added this year, you could average my results from this year plus the last two years). A second is an analytical approach, which brings us to uncertainties about how we should estimate growth.

One of the biggest failures of education policy in recent years has been the failure to adequately evaluate the accountability systems that were imposed on teachers and students. The movement toward value-added models exacerbates this because of serious gaps in our knowledge of their workings and effects.

4. How certain are we about how to model gains?

A variety of different statistical approaches are used to estimate value added. Most are highly complex, and while the differences among them seem extremely arcane, in this case, the old cliché really is true: the devil is in the details. The choice among methods can matter; it can influence, sometimes substantially, how a school or teacher is rated. And yet, other than the experts, few people understand how these models work or what the implications of the various choices are. Let's look at a handful of the more important technical issues.

One important issue is how to deal with the uncertainty caused by sampling error. All teachers will sometimes appear more or less effective than they really are because of sampling error, and substantially incorrect estimates will be much more common among teachers with smaller classes (or schools with smaller enrollments). One approach ignores the fact that these errors are worse in small groups and takes each group's estimate at face value. The alternative approach, called a "random effects model," compensates for the uncertainty by "shrinking" the estimates for each teacher or school back toward the average teacher or school, with more shrinkage for the groups with fewer students. In the aggregate, the latter approach seems preferable, because it compensates for small samples, puts large and small groups on the same footing, and reduces the number of instances in which a teacher or school is inappropriately rewarded and sanctioned because of sampling error. For individual teachers or schools, however, this approach is not necessarily fair. For example, if you happen to be an exceptionally effective teacher but have a small class, a random effects model will assume that the atypically rapid growth of your students reflects sampling error and will shrink it. Therefore, random effects models reduce one type of error but increase another: the probability of missing truly effective or truly ineffective teachers.

Another issue pertains to the persistence of the effects of teachers. Value-added models ask the question: how much has the year with you added to students' growth *given what prior experience contributed*? To answer that question, one first has to estimate those prior contributions, and the different ways in which various VAMs do this can affect how teachers are rated. Suppose you receive a group of students who had highly effective teachers the previous two years, and suppose that the students score very well at the end of your year with them too. To calculate your value added, one has to somehow subtract what the students would have known at the end of your year, given their prior experience. The more the effects of that prior good teaching persist, the less credit you deserve for the students' strong performance at the end of the year. One of the most common models, the "layered model," assumes that the impact of good or bad teaching persists forever without any lessening at all. (As a teacher, I find this hard to accept; I could only wish that everything my students learned persisted without any deterioration.) Other models, however, allow for an erosion of prior teachers' effects over time, giving you more credit (or blame) for the performance of students at the end of their year with you. Decisions about how to handle persistence can clearly influence how individual teachers or schools are rated.

Another choice is how to deal with missing data. All value-added models require longitudinal data, that is, data that track individual students over time. However, some students—and in some districts or schools, many students—do not have complete data. Their data may be missing for all manner of reasons: their

families moved, they were truant, they were assigned to a special class, and so on. What is important is that the students whose data are missing are often unlike those whose data are complete. Worse, we generally know only enough to discern that these students are different; we do not know enough about them to adjust for the effects of leaving them out of the calculation. Some of the VAMs can handle missing data, provided that the problem is not too severe, but it remains an open argument just how serious this problem has to be before it substantially biases estimates for some teachers.

Apart from the first of these issues, all of these are matters of bias, not error. For example, if we overestimate persistence, we will introduce a bias by systematically over- or underestimating the impact of teachers depending on the effectiveness of those who preceded them.

5. How well do we adjust for other influences on achievement growth?

To provide an unbiased estimate of the effects of teaching, value-added models must remove the impact of other influences on achievement growth. Teachers often express concern that the models now used will not do this well enough to be fair. For example, many teachers find that their effectiveness varies with the characteristics of their students. I certainly have; my style and methods of teaching work much better with some types of students than with others. If these effects are large, value-added models would have to take them into account.

Teachers are right to be concerned. On the positive side, a recent study* found that in one context, effectiveness as estimated by a value-added model was similar to true effectiveness measured by an experiment, but there are a number of reasons why we cannot in general assume that this is true.

One potentially important source of bias in the evaluation of teachers is called "interference." Suppose you want to evaluate the impact of providing after-school math tutoring, and you do this by randomly dividing students from a school into two groups and giving tutoring to only one of them. You give both groups a math test at the end of the year, and you use the difference in scores between the groups to evaluate the impact of the tutoring. This sounds like an ideal evaluation—a true experiment. The problem is that the tutored and untutored students interact with each other: they attend the same math classes, they may study together, and so on. This is interference: the effects of tutoring seep into the untutored control group, leading to a biased estimate (in this case, too low) of the impact of tutoring. Interference is a potentially severe problem in using VAMs to evaluate teachers because teachers are embedded in schools, and there are many sources of interference that could bias estimates for individual teachers. Interference could arise not just because of the instruction of other teachers, but because of administrative arrangements, peer effects, and so on. For example, in some secondary schools, teachers in subjects other than math and English have been instructed to incorporate more math and writing into their classes, which makes the value seemingly added by math and English teachers dependent in part on the other teachers their students are assigned to. For this reason, some researchers have warned that with the value-added models we have now, the effects of teachers cannot be entirely separated from those of the school context.

Apart from interference, there is an ongoing, intense debate about how well VAMs control for other factors that influence achievement growth, such as students' backgrounds. The adequacy of the models is likely to vary, depending on the context (for example, the degree to which students with similar characteristics attend the same classrooms and schools) and the methods used. The more similar the contexts in which two teachers work, the less these other factors come into play, and the closer a value-added model will come to an estimate of the teachers' impact. But in real-world situations in which the contexts of teaching vary markedly (even within a single school), research tells us that we can't assume that the results of our models give us a sufficiently unbiased estimate of the effects of teaching.

6. How does score inflation affect value-added models?

In this era of test-based accountability, one of the biggest problems confronting testing programs is *score inflation*: increases in test scores that are larger than the actual gains in learning they are thought to represent. Research has shown that score inflation is widespread and that it can be very large. Some studies have found score gains that are three to five times as large as they should be, and others have found large score gains that were not accompanied by any meaningful improvements at all. Score inflation results in both an illusion of progress and misleading comparisons of schools and teachers, both of which are detrimental to students. (A more detailed explanation of score inflation, as well as a discussion of the grey area between good instruction and inappropriate test prep, are included in the sidebar from *Measuring Up* on page 22.)

VAMs do nothing to address the problem of score inflation. There may be ways that policymakers can lessen this problem, such as relying on multiple measures, setting more realistic targets, and strengthening the role of human judgment in the evaluation of teachers and schools, but simply switching from status or cohort-to-cohort change models to a value-added approach will not do the trick.

Where Do We Go from Here?

For all the uncertainties and concerns about the use of value-added models, there is no question that they are in some important ways superior to the status and cohort-to-cohort change models that have dominated test-based accountability in the United States for the past 30 years. I believe that most people working in this area would agree with me that we should continue to look for appropriate ways to incorporate value-added modeling into accountability systems in order to capitalize on that superiority.

At the same time, to use value-added models sensibly, we can't treat them as a silver bullet. We need to find ways to use VAMs that take into account both their limitations and the uncertainties we still have about their functioning and impact.

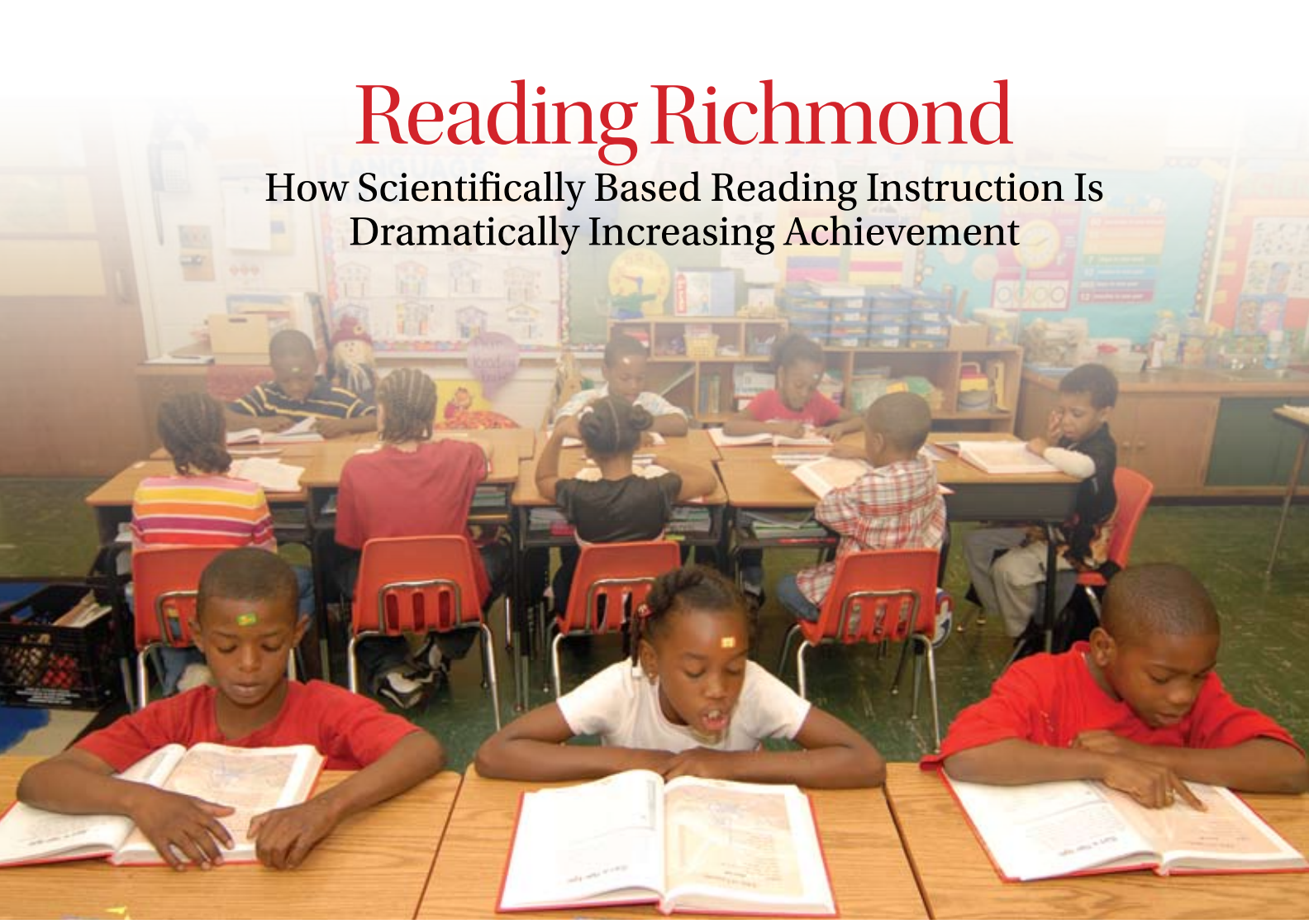
First, we must recognize that value-added modeling remains

(Continued on page 39)

* S. Cantrell, J. Fullerton, T. J. Kane, and D. O. Staiger, *National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment* (National Board for Professional Teaching Standards, June 11, 2008, draft).

Reading Richmond

How Scientifically Based Reading Instruction Is Dramatically Increasing Achievement



By Jennifer Dubin

It was a typical day in Kimberly Bailey's second-grade classroom. Her students played in a sleeping bear's cave, made friends with animals named Badger, Mouse, and Gopher, and attended a small party in their honor. No guest from the local zoo walked around the room. No special visitor held their attention. Yet the students, clearly excited, constantly raised their hands to participate in the class discussion. So what accounted for their enthusiasm? Something as simple as reading a book aloud to each other.

But not just any book. The textbook these students were reading is *Hiding Places*. As its title suggests, the book features reading passages about animals and their habitats. It's specifically geared toward second graders and is part of a scientifically based reading program.

There is indeed a science to teaching children how to read. In 2000, the National Reading Panel issued a report based on a comprehensive review of reading studies. The panel found that early reading instruction ought to include explicit teaching of five key components: phonemic awareness (identifying and being able to manipulate the sounds in words), phonics (understanding how letters are linked to sounds), fluency (reading orally with speed, accuracy, and proper expression), vocabulary (understanding the meaning of words), and text comprehension (understanding whole passages). Instruction that focuses on these five components is especially important for children who have had little to no exposure to print before they begin school. And, according to Bailey, this type of instruction is exactly what the children in her class need.

Her students attend Fairfield Court Elementary School in Richmond, Virginia. The school, like the district, is majority African American. And like the district, its students mostly come from low-income families. Of the roughly 500 students enrolled in the school, 97 percent receive free or reduced-price meals. That's 26 percentage points higher than the district and 64 percentage points higher than the state.

Fairfield Court is in Richmond's East End, which has high rates of poverty and crime. Despite such challenges, an important story about student achievement there, and across the entire city, has begun to emerge. Since Richmond Public Schools started to focus on research-based reading instruction eight years ago, the reading scores of its students on state assessments have climbed substantially. (See the charts with third- and fifth-grade results, the only elementary grades with longitudinal data, on page 32.)

Jennifer Dubin is assistant editor of American Educator. *Previously, she was a journalist with the* Chronicle of Higher Education. *Photos by Michael Campbell.*

Of course, reading programs alone did not raise achievement in the district. The schools benefited from a new superintendent, an overhaul of the central office, and more support for more targeted approaches to professional development. As many teachers in Richmond will quickly tell you, programs don't teach reading; teachers do.

At the same time, educators like Jean Gritz, a first-grade teacher at Fairfield Court, readily attest to the effectiveness of research-based reading programs—how phonemic awareness, phonics, fluency, vocabulary, and comprehension instruction have helped them reach their students. "We tell the children," says Gritz, 'If you read, you can do anything.'"

Kimberly Bailey, a second-grade teacher at Fairfield Court, spends the first 45 minutes of each two-hour reading block teaching students in a large-group lesson. Using a research-based reading program that the district began implementing eight years ago, she focuses on phonemic awareness, phonics, fluency, vocabulary, and text comprehension. Instruction that focuses on these five components is especially important for children who have had little exposure to print before they begin school.

The Need for a Phonics-Based Program

Richmond's success in reading did not happen overnight. First, administrators had to figure out what the district was doing wrong. In 1999, Yvonne Brandon, who is currently serving as the district's interim superintendent, had just been appointed the director of instruction when she was charged with unpacking students' low test scores. "One of my first tasks was to find out just what we were using in areas of reading, especially elementary." She surveyed the schools and found that at least 29 different reading programs were being used. Programs varied from school to school, even within schools.

Having a coherent curriculum is crucial for districts like Richmond with high student mobility. Richmond's mobility rate is more than 40 percent. All that variation in the reading programs hampered student achievement, since many children would start the year in one school, and then have to adjust to a different program each time they moved. But Brandon noticed the district did have one program that seemed to work well: a Voyager reading series used in elementary summer school. Called Time Warp, it took kids on a journey through history. "We saw great gains," Brandon says, because the program was meeting students' needs. "The data showed we needed a program strongly based in phonics."

At the time, Voyager published only the summer program, which the district continues to use in summer school as an intensive intervention for students who are behind. But in 2000, the company created a year-round program for grades K-2,* the Voyager Universal Literacy System, and Brandon traveled to Voyager's company headquarters in Dallas to see it. She recalls being impressed by what she found.

The program has a different adventure theme (such as sea castles or hiding places) for each grade that is designed to increase students' reading skill, vocabulary, and background knowledge by having a mix of fiction and nonfiction texts. For instance, the first-grade reading program includes *Hercules the Harbor Tug*, a story about boats with pictures and passages that familiarize students with words such as buoy, channel, and dock.

Teachers in each grade receive a detailed manual complete with lesson plans for a daily two-hour literacy block that includes a 45-minute large-group lesson, 60 minutes for reading stations, and then a 15-minute writing, vocabulary, or spelling lesson. For the reading stations, teachers place students in three groups, which rotate every 20 minutes. Students work together at two of the stations on recently introduced reading skills. At the third station, students work with the classroom teacher, who follows a detailed lesson plan to give students small-group instruction.

At the beginning of the year, students take assessments to determine whether they are "struggling," "emerging," or "on-track" in key literacy skills like letter-naming fluency for kindergartners or reading connected text for second graders. These assessments are equivalent to the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a set of standardized, individually administered one-minute measures of early literacy development.†

Students who score at the "struggling" level receive additional instruction during the day and take weekly progress-monitoring assessments until they master the skills in question. There's also an Extended Time Curriculum for "struggling" first and second graders, which reinforces the reading skills they are learning during the regular literacy block.

Students who score at the "emerging" level are also carefully

* This program has since been expanded to include third grade, but Richmond still uses it only in grades K-2.

† To learn more about DIBELS, see "Preventing Early Reading Failure" by Joseph K. Torgesen in the Fall 2004 issue of *American Educator*, available online at **www.aft.org/pubs-reports/american_educator/issues/fall04/reading.htm**.

Above, at Fairfield Court, students have a 45-minute enrichment class every day. That provides time for regular classroom teachers to meet in grade-level teams and ensures that the children have time for art, music, P.E., and media classes.

Below, Joyce Williams, a fourth-grade teacher at Fairfield Court, works with a student during the literacy block. A decade ago, Richmond's elementary schools were using at least 29 different reading programs. Today, they are using just two, both of which are research based.

monitored and receive targeted instruction. They are assessed once a month until they reach the "on-track" level.

Students identified as "on-track" do not take weekly or monthly assessments. They, as well as all students, take a set of one-minute benchmark assessments three times a year.*

Monitoring students' progress and delivering targeted instruction is demanding, so teachers also receive intensive professional development. When a school or district first adopts the program, a trainer from Voyager provides a two-day training for district and school-based "coaches" (usually Title I reading specialists) and a three-day training for teachers. Then, spread across the school year, there are eight three-hour monthly training sessions that consist of teachers practicing direct instruction, administering assessments, grouping students, and modeling lessons. There's also ongoing professional development throughout the year delivered by the coaches. They visit classrooms and model lessons to help teachers hone their instruction. They also help teachers use student data to inform their instruction.

With so many teaching materials, embedded assessments, and significant amounts of embedded professional development, Brandon liked the Voyager program immediately. "I came back excited," she says.

Overcoming Doubt

In 2000, Brandon conducted focus groups with teachers and principals who attended Voyager demonstration lessons held at an elementary school in the district. Not everyone shared her enthusiasm. She recalls that some veteran teachers, used to relying solely on a single textbook, thought the program offered too many instructional tools. Nonetheless, Brandon persuaded the district's top administrators to pilot the program in 2000 in a handful of schools with very low reading scores. Then, in 2001, the district added a few more low-scoring schools, including Fairfield Court.

"That first year, I'll never forget," says Velicia Coleman, Fairfield Court's Voyager coach and Title I reading specialist. "There were reluctant teachers. They were coming from a program where they had complete control, and they could do what they wanted." So the Voyager program, which has a detailed teacher manual, was not always well received. Some teachers objected because they couldn't keep up with the time limits for delivering whole-class and small-group instruction. And they didn't like timing their students on one-minute reading tests. Teachers would say of a student who didn't pass the tests, "I know he knows it, but he didn't do it in one minute," Coleman recalls.

She remembers her own uncertainty as to whether such short assessments could measure a student's reading ability. "How in the world can you project what a child can read after one minute?" she recalls thinking. After using the assessments, she realized they worked. "You can tell if a child is on track or not, and you can find out immediately." One-minute assessments work because in reading, efficiency (or "automaticity") is important. Although children initially become accurate readers by learning to decode words through phonics, they must eventually learn to recognize most words instantly in order to become fluent readers.

Once teachers began to follow the program, they saw results with their students. Those results, though, didn't materialize just because teachers followed the manual; they materialized

* This is one way in which Richmond differs from the official Voyager program. Richmond's students take benchmark assessments four times a year. The initial one assesses students' reading ability; the others monitor progress throughout the year.

because teachers put their personalities into the program. "You have to have a little bit of gusto to do a Voyager lesson," Coleman says. "You can't just get up there and read a statement with no expression. If you put a little life in it, the kids are going to listen."

Coleman, herself, initially doubted the program. A reading specialist since 1988, she had seen her share of educational fads. Then one day in the spring of the program's first year, she observed a kindergarten class at the school. She recalls, "The kids kept saying, 'Ms. Coleman, I want you to hear me read.' I stayed and I listened and I was amazed." The students read much better than she had ever heard kindergarteners read. She remembers she wore white pants that day. After leaving the classroom, "I had all these handprints all over my pants because the kids were eager to show me they could read." The experience convinced Coleman that the program would work.

Second-grade teacher Kimberly Bailey has decorated a "Word Wall" with letters of the alphabet and words that begin with each letter. Students use the wall as a reference during independent and group work in class.

Signs of Improvement

Resistance to trying a new approach to reading instruction districtwide did not diminish until Richmond reached a low point. In 2001-02, Brandon recalls, "We were declared the second lowest school division in the state of Virginia." That was "a point of embarrassment." Only then did teachers and administrators agree it was time to make some big changes.

In 2002, Deborah Jewell-Sherman became Richmond's superintendent, and she made sure that when it came to scientifically based reading instruction, everyone was on board—but she didn't force all of the elementary schools to adopt Voyager.† Instead, in 2003, the district piloted Houghton Mifflin Reading, another research-based program, in eight elementary schools. Yvonne Brandon says district officials were drawn to it because, like Voyager, it focused on the National Reading Panel's five components of reading instruction, and it offered extensive professional development, embedded assessments to monitor students' progress, and plenty of work on comprehension and writing. For instance, the program features weekly teacher read-alouds, in which students listen to the teacher read aloud a particular passage and then respond to a series of questions. The read-alouds help students expand their vocabularies and improve their comprehension. Also, in grades 3 through 5, books in the program's "Reader's Library" continue to reinforce high-frequency vocabulary words. Under Jewell-Sherman's watch, in 2003 all elementary schools in the district implemented one or both of these research-based reading programs. Today, 10 of the district's 28 elementary schools use Houghton Mifflin for kindergarten through fifth grade, and 18 elementary schools use Voyager for kindergarten through second grade and Houghton Mifflin for third through fifth grade.

To facilitate the adoption of research-based reading instruction, Richmond also applied for, and won, a Reading First grant. (Reading First is a federal program that supports the implementation of research-based reading instruction; see sidebar, page 34.) Today, five elementary schools receive Reading First grants and, to extend the program's reach, the district has created a Reading First consortium. The consortium consists of 15 elementary schools, five of which receive Reading First awards and 10 others (including Fairfield Court) with test scores that signaled they needed more district support. The consortium includes the principals and reading coaches of these 15 schools. They meet monthly with Victoria Oakley, the district's director of instruction, to discuss the five components of reading instruction, how reading permeates all subject areas, and what to look for during class observations. Each semester, the group selects a book to read for professional development. Last spring's topic was fluency; the group read *The Fluent Reader: Oral Reading Strategies for Building Word Recognition, Fluency, and Comprehension* by Timothy V. Rasinski.

Over the past several years, schools in the Reading First consortium also benefited from other kinds of intensive district support. For instance, five years ago, instructional specialists from the central office often visited these schools monthly. Because of the schools' improvement, specialists now visit them every nine weeks, but they are available at the principals' request.

Benefits continue to extend across the district, too. For example, the lessons learned about the need for ongoing professional development are now being applied districtwide, and not just in reading. "We used to have all teachers come to huge professional development sessions," Brandon says. Teachers in the same grade level and in the same subject would meet on an in-service day in whatever high school could hold them. That set-up "wasn't providing them with the intensive training they needed." Now department heads and lead teachers hold professional development sessions in their own schools, a more targeted approach.

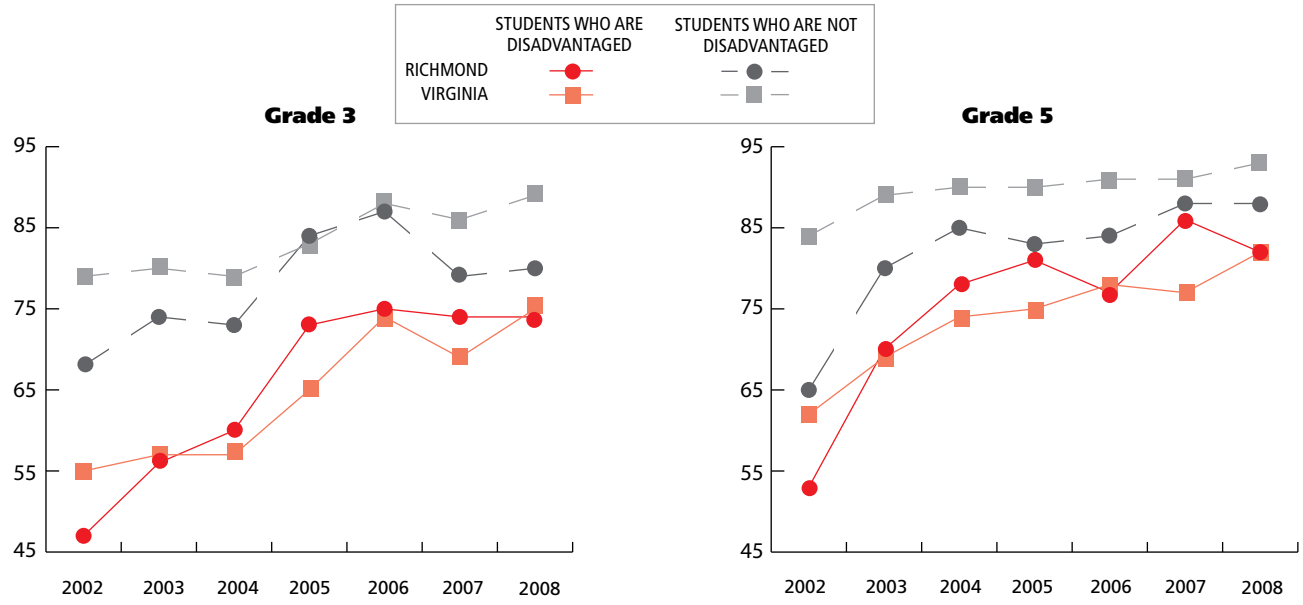
Since these changes, passing rates on state reading assessments have jumped. For instance, in 2001–02, 53 percent of economically disadvantaged fifth graders passed. By 2007–08, 82 percent did. Even better, student achievement gains in the district have extended beyond reading, resulting in dramatically more elementary schools being fully accredited. In 2002–03,

† Jewell-Sherman resigned as superintendent in July and is now at Harvard Graduate School of Education.

Reading Achievement Soars for Richmond's Disadvantaged Students

Since Richmond Public Schools began to implement research-based reading programs in its elementary schools eight years ago, reading achievement has increased substantially. The charts below show that Richmond's economically disadvantaged third and fifth graders (the only elementary grades with data going back to the 2001-02 school year) are now passing Virginia's reading tests at rates as high as their state counterparts. This is no small feat considering that Richmond's poverty rate is more than double the state's: 71 percent of Richmond's students, compared with just 33 percent of students statewide, are eligible for free or reduced-price meals. Richmond's third and fifth graders who are not economically disadvantaged have also made important gains; they are now passing at rates almost as high as their state counterparts.

Percentage of Students in Richmond and in Virginia Who Passed the State Reading Assessment, Broken Down by Those Who Are and Are Not Economically Disadvantaged



Note: The data presented here on students who are and are not "disadvantaged" were drawn, in September 2008, from the Virginia Department of Education's online Virginia Assessment Results database, available at **https://p1pe.doe.virginia.gov/datareports/assess_test_result.do**. In determining who fits into its "Students Identified as Disadvantaged" subgroup, Virginia has several criteria, such as eligibility for free or reduced-price meals, eligibility for Medicaid, or homelessness.

7 of the district's 29 elementary schools were fully accredited. In 2007–08, 26 of the district's 28 elementary schools were fully accredited.*

Learning to Read

A visit to Kimberly Bailey's class at Fairfield Court reveals the story behind the numbers. One morning in April, during the two-hour literacy block, Bailey reviews with her 17 second graders a book she had read to them the day before: *Bear Snores On*. The book, part of the Voyager program, is specifically designed for grade 2. A quick flip through its pages reveals colorful pictures and language full of repetition and rhyme: "In a cave in the woods, in his deep, dark lair, through the long, cold winter sleeps a great brown bear. Cuddled in a heap, with his eyes shut tight, he sleeps through the day, he sleeps through the night. The cold winds howl and the night sounds growl. But the bear snores on."

Bailey stands at the front of the room and writes "Bear" on the board. When she asks why the word "bear" is sometimes capitalized in the story, a student says because it's the animal's name (meaning that it's the character's proper name). Bailey then asks students to give the names of some of the story's other characters. Little voices call out "Badger" and "Raven."

"Cheyenne, what family is Raven in?" Bailey asks.

"A bird family," Cheyenne says.

Bailey then asks students to define setting (when and where the story takes place, they answer) and what this story's setting is (in a cave at night, they say). She jogs their memory about the book and writes what happened on the board: a "small fleck of pepper made the bear sneeze." After jotting down some more story details, she tells the students the notes on the board are "our background information."

Next, Bailey turns on the overhead projector and tells students

* Elementary schools are fully accredited if: (1) they have a combined pass rate of at least 75 percent on English tests (which include reading tests) in third through fifth grades; (2) they achieve pass rates of at least 70 percent in mathematics in third through fifth grades and in fifth-grade science and fifth-grade history; and (3) they achieve pass rates of at least 50 percent in third-grade science and third-grade history. (Source: **http://www.doe.virginia.gov/VDOE/src/accred-descriptions.shtml**.)

they have two minutes to edit two sentences. The first one reads, "Do bares really snore when sleep?" One girl gazes up at the bulletin board across from her for possible clues. The board is a "Word Wall" that Bailey has decorated with letters of the alphabet and words that begin with each letter. Next to "Aa" is "about, after, again." Next to "Jj" is "joke, jump, junk." When Bailey calls time, a student named Trenajah says, "Bears is spelled wrong." Bailey asks for the correct spelling and the class calls out "b-e-a-r-s." After she edits the sentence, she asks if she can change anything else. A boy says the sentence needs a period. Bailey asks if somebody can tell her why the sentence doesn't need one. "Because we're asking a question," a student says. Seconds later, Cheyenne tells Bailey, "You need to put 'they' between 'when' and 'sleep.' " The students then agree their editing is complete. They have correctly spelled and punctuated sentence number one, and go through a similar process for sentence two.

At Fairfield Court, there is no narrowing of the curriculum. Students take reading, math, science, and social studies daily, and they have art, music, P.E., and media classes each week. Here, a student proudly displays her art work.

A few minutes later, Bailey allows them 15 minutes to make props for a play they will perform in class that day based on *Bear Snores On*. They paste brown and green paper for trees on white construction paper. And they color in a narrow band of blue sky at the top of the paper, the way kids normally do. A semicircle cut out of a grocery bag serves as the focal point: the bear's cave. Bailey helps them put their small props underneath the board at the front of the room. "Look at what you came up with in 15 minutes!" she laughs, delighted with their work. Bailey's enthusiasm and energy are infectious and certainly reflected in her students' eagerness to participate in class.

After assigning parts, the students, holding pictures of their characters glued to popsicle sticks, read aloud the play, "Party Time!" from their books. The play is based on *Bear Snores On* and includes the same vocabulary and characters. Bailey asks them to repeat words or sentences when they make mistakes.

After the play, she reads aloud a short story the class wrote for the city's upcoming literary festival. Then she asks the students to gather in their groups and work in stations. One group plays a spelling game that reinforces some letter patterns the class has been learning, another group does an exercise from their books asking students to write the sequence of events in the play they performed, and the other group Bailey asks to rewrite the ending of *Bear Snores On* any way they wish.

A quick look at her teacher's manual, which she always keeps close by, reveals that Bailey followed the morning's lesson to a tee. Yet, she clearly made the lesson her own and was energized by teaching her students how to read.

But if Bailey had experienced any trouble, help would not have been far away. Velicia Coleman's job, as the school's Voyager coach and Title I reading specialist, is to ensure that every teacher gets what she needs. Three times a week, Coleman conducts classroom observations, what she calls "walk-ins," where she makes sure teachers have the support they need. Last year, for example, she worked with a teacher who had trouble delivering instruction. During classroom observations, Coleman noticed that the teacher had not grouped her students into different work stations and that she was teaching some letter combinations and the sounds they represent incorrectly. To help her improve, Coleman and the school's principal held a conference with the teacher; Coleman also modeled lessons for the teacher and gave her one-on-one support. When she returned to observe the teacher's classroom a week and a half later, Coleman says "the improvement was there."

Coleman also fills in as needed. If a regular classroom teacher is out one day and the substitute has not been trained in Voyager, Coleman teaches the literacy block herself. And, Coleman does remediation for students who need extra support. All these roles make for a full schedule, but they also help ensure that students receive consistent instruction—a big improvement over the multiple programs and instructional approaches once common in Richmond, even within individual schools.

Jean Gritz, who has taught first grade at Fairfield Court for 30 years, appreciates the supports embedded in a research-based program. She likes the continuity, the repetition, and the time built in for review, all of which allow children who don't get something the first time to pick it up the next time. As a result, her students can pretty much read on their own by midyear. Before, they couldn't do so until March or April. "If you do the program as it's designed to be done, I can't see you failing," she says.

Keeping the Curriculum Broad and Rich

While teachers intensely focus on helping students learn to read, literacy instruction in the district does not happen at the expense of everything else. There is no narrowing of the curriculum—a fact that contributes to students' success.

The typical school day at Fairfield Court consists of a two-hour literacy block, then a 90-minute math block, a one-hour block of science, and 45 minutes of social studies. Students still get

recess every day, as well as art, music, and P.E., which they attend on a rotating basis Monday through Thursday. On Fridays, students have class in the library for a media lesson. During these 45-minute enrichment classes, regular classroom teachers get time for planning lessons together by grade level.

On Tuesdays, Wednesdays, and Thursdays, students can stay after school until 5:15 p.m. for an extended day. They can practice their reading skills on the computer, play other enrichment games, or do their homework. Roughly 150 students stay after school each of those three days. On Saturdays, from 9 a.m. to 12 p.m., anywhere from 50 to 70 students attend the school's Saturday Academy, where students focus on reading and math. (Teachers and staff say they try to keep the students in school as much as possible to give them a safe haven.)

From June 23 through July 28, there's also summer school from 9 a.m. to 2 p.m. Students identified as struggling readers are invited to enroll so they can improve their reading skills. The summer school curriculum still features Voyager's Time Warp, which takes students on a theme-based trip through history. For instance, second graders study ancient Egypt. At Fairfield Court this summer, roughly 85 students were enrolled in summer school in kindergarten through fifth grade. The summer school also has an extended day on Tuesdays, Wednesdays, and Thursdays until 5 p.m. About 50 students stayed after school each of those days.

Although scores at Fairfield Court have risen in the last few years, teachers there continue to face challenges. In the fall of 2006, 225 students from Whitcomb Court Elementary School transferred to Fairfield Court after their school closed because of declining enrollment. "That has accounted for a lot of our discipline problems," says Irene Williams, Fairfield Court's principal. The number of incidents of disruptive behavior skyrocketed from 63 in 2005-06 to 1,360 in 2006-07 (the most recent year for which figures are available). Williams attributes the increase to the new students adjusting to the school.

Although figures for disruptive behavior for the 2007-08 academic year are not yet available, school officials believe the situation has improved. And yet, even with all the behavior challenges, Fairfield Court students have continued to succeed

(Continued on page 36)

Does Reading First Deserve a Second Chance?

Reading First is a federal program designed to support schools in implementing research-based reading instruction. It has come under fire recently and, as American Educator *goes to press, its future funding is uncertain. Some of the controversy is due to allegations of mismanagement and some is due to claims that the program is not very effective. However, researchers from the Northwest Regional Educational Laboratory have found that Reading First is having a positive impact—and that impact may even extend to schools without Reading First grants. Here's a brief summary of their four-year evaluation and of the concerns they have with claims that Reading First isn't working.*

–EDITORS

BY THERESA DEUSSEN, KARI NELSESTUEN, AND CAITLIN SCOTT

Since 2003, Reading First has provided unprecedented amounts of federal funding to states for K–3 reading programs, with the goal of having children read at grade level by the end of third grade. Reading First, however, is more than just a funding source. Schools awarded grants were required by federal legislation to use curricula and practices that were grounded in "scientifically based reading research." These included using a research-based core reading program, hiring a reading coach, providing at least 90 minutes of reading instruction per day, assessing students' reading skills regularly, and providing reading interventions to struggling students. States were responsible for providing grantee districts and schools with the professional development and technical assistance necessary to implement these and other Reading First requirements.

Each state was also required to hire an independent organization to conduct an annual evaluation. Our organization, the Northwest Regional Educational Laboratory (NWREL), was hired as the external evaluator in four states, Alaska, Montana, Washington, and Wyoming, and it also contributed to the evaluation in a fifth state, Arizona, in collaboration with the Arizona Prevention Resource Center at Arizona State University. The evaluation in each state examined Reading First implementation as well as student achievement outcomes. These evaluations were designed to help states make ongoing, data-based decisions about their program.

As researchers, we know a single study is never able to capture all the information that can be gained about a particular program or initiative. Instead, it takes multiple studies over time to provide a rich and accurate understanding of how well a program works. This is why the oversimplification of findings from the recent interim report of the federally funded Reading First Impact Study is troubling. That study found no significant differences in performance on a comprehension measure between students at a subset of Reading First schools and students at non-Reading First schools in the same districts.[1] Some media coverage interpreted this finding simply as "Reading First doesn't work."[2]

The findings of the impact study are important, but they do not tell the entire story. NWREL's statewide evaluations of

Theresa Deussen is unit director for Language and Literacy Evaluations in NWREL's Center for Research, Evaluation, and Assessment, where Kari Nelsestuen is senior advisor and Caitlin Scott is evaluation advisor. *This article is adapted from "Does Reading First Work? Data Trends from Evaluations in Five Western States," published by NWREL in June 2008 and available online at* ***www.nwrel.org/crea/pdf/rf-trends.pdf****.*

Left and center, Fairfield Court's focus on building students' literacy does not come at the expense of everything else. Students still enjoy 15 minutes of recess every day. Below, when students are in class, they are engaged.

Reading First provide a more nuanced picture of the program. Across the five states, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) was NWREL's primary measure of student outcomes. This assessment includes a set of standardized, individually administered measures of early literacy development. Students obtaining adequate scores on these assessments are said to be "at benchmark," while the students scoring at the lowest level fall into what is commonly called the "intensive group."

On the DIBELS assessment, NWREL's statewide evaluations found that there was steady improvement in the percentage of students performing "at benchmark," and a decrease over time in the percentage of students performing at the lowest ("intensive") level.

In addition, NWREL's evaluation of the implementation of Reading First revealed a trend that raises some questions about the validity of comparing Reading First and non-Reading First schools within the same district, as the impact study did. Across the five states, the evaluations found that in districts with Reading First grants, non-Reading First schools frequently implemented many Reading First program components. Survey data from the five states showed that many non-Reading First schools routinely used other funding sources (most often district funds) to implement key components of Reading First, such as a scientifically based core reading program, a reading coach, regular assessments, and systematic interventions for struggling students.

These results suggest that Reading First has had an impact that extends beyond the schools directly receiving grants. This "spillover" complicates any comparison of Reading First schools with non-Reading First schools since, in essence, many non-Reading First schools implemented similar reading programs. It may be that the impact study did not find differences in student achievement because the non-Reading First schools were implementing many of the components of Reading First.

Like the national impact study, NWREL's evaluations had their own limitations, most importantly the lack of comparison groups and the fact that DIBELS does not measure comprehension. Still, the consistency of findings across states and over time is suggestive of positive impact.

Reading First is a complex, multifaceted program implemented in many different school and district contexts across the country. It is not surprising that multiple evaluations should come to different conclusions about both implementation and outcomes. These variations make it all the more crucial that policymakers and practitioners consider multiple reports and data sources (and their limitations) before making decisions that will affect the education of many thousands of disadvantaged students in some of the poorest schools in the nation.

Endnotes

1. Beth C. Gamse, Howard S. Bloom, James J. Kemple, and Robin Tepper Jacob "Reading First Impact Study: Interim Report," NCEE 2008-4016 (Washington, DC: U.S. Department of Education, April 2008).
2. Nancy Zuckerbrod, "Study: Bush Administration's Reading Program Hasn't Helped," *USA Today*, May 1, 2008; and Kathleen Kennedy Manzo, "Reading First Doesn't Help Pupils 'Get It,'" *Education Week*, May 7, 2008.

Reading Richmond

(Continued from page 34)

academically, thanks in part to research-based reading instruction, and according to the principal, devoted teachers. Because such programs—and the ongoing support that teachers have received to implement them—have worked well in the district, it appears they are here to stay. Increases in achievement will ensure that, says Yvonne Brandon, the district's interim superintendent. "What excites me now is to go to a class and see the kids clamoring to get certain book titles because they know what the book brings to them," she says. "They can escape from whatever is going on around them. They go into a world of language."

That world can differ strikingly from their own. In Joyce Williams' fourth-grade class at Fairfield Court one morning in April, students discuss two sports, cricket and baseball, after reading a brief passage. Cricket, they learn, originated in England and lasts from one to four days, while baseball has nine innings that take just one afternoon. Both sports, though, are played with bats and balls. Williams uses the passage to explain the terms "compare" and "contrast." Besides allowing them to practice comprehension skills, the passage helps students acquire new vocabulary, ponder life in a foreign country, and learn about a sport they don't play at home.

Their teachers' hope, though, is that another, more important lesson will begin to sink in: the more you read, the more you know. □

Peer Assistance and Review

(Continued from page 11)

allowed me to say you pick your battles and to be honest, you know, it's phenomenally hard to get rid of somebody. So I would say, 'Do I want to take the time to [get rid of them], knowing that I've also got this, I've got that, etc.' So you say, 'No.' "

Solution: The California legislation included teachers' unions as partners with districts in a couple of ways. The legislation required that the union sign off on a district's proposal to the state creating a PAR program (and it is worth noting that the district would lose state money it was already receiving if it did not create a PAR program). In addition, the legislation required that the panel be co-led by the union and the district, and that it be made up of five teachers and four administrators. In these ways, the Rosemont teachers' union played a central role in the changes brought about with PAR. The survey results indicate that principals, panel members, and consulting teachers all thought PAR had a positive effect on relations between the teachers' union and the district. One principal highlighted the change: "I'm working collaboratively with the union. It's a whole different feel and there's a sense that the union and I agree that we need teachers who use best practice, and we're working together to have best practices occur, and we're not opposed in terms of keeping some person in there who is not utilizing best practice. I feel like we're all on the same team and it's about children and the kind of teaching they get." Some principals were quite surprised to see the teachers' union president sitting at the table at hearings, let alone arguing for dismissals of teachers. PAR programs, however, have historically been initiated by union presidents interested in "postindustrial unionism,"[25] and it was the union president who advocated for the creation of a PAR program in Rosemont prior to the implementation of the state legislation. For some teachers' unions, PAR is one way to defend the profession of teaching rather than individual teachers.[26]

5. Generating Confidence in Evaluative Decisions

Problem: Principals often doubt themselves when making evaluative decisions.[27] How could it be otherwise? The problem of making a decision has accrued through the problems discussed above. Principals do not have sufficient time to spend on evaluations and are not involved in professional development in an ongoing and substantive manner; therefore, they are uncertain that the teacher under review has been given an opportunity to improve. They typically lack standards on which to rate teachers. They are alone to make the decision, without the benefit of an organizational structure that provides collaboration with colleagues. Finally, they often believe that a negative evaluation of a tenured teacher will involve a timely and costly battle with the teachers' union and that they will likely lose that battle.

Solution: Just as the problem of making a decision accrues through the prior problems, so the solution accrues through the prior solutions. Due at least in part to the amount of time devoted to assisting the participating teachers, the ongoing nature of the reviews, the link between the reviews and teaching standards, and the shift from one reviewer standing alone to a group of peers participating in the process, consulting teachers, princi-

pals, and the panel had an increased sense of confidence in the quality and accuracy of the reviews. While my study did not examine the teacher evaluation paperwork, people involved in PAR, including principals, believed that higher-quality evaluations were being conducted through PAR than had occurred through the traditional process.

6. Increasing Accountability for Teaching Quality

Problem: Given the structural weaknesses in the traditional system of evaluation outlined above, teachers rarely are fired for teaching poorly.[28] In one study of traditional teacher evaluation, less than 1 percent of teachers were dismissed, despite the fact that 1.53 to 2.65 percent were formally identified as "incompetent" and 5 percent were informally identified as "incompetent."[29] Such teachers are more likely to be reassigned to other school sites than fired.[30]

Solution: Perhaps one of the most significant findings in the study is that, across the board, consulting teachers were willing to recommend nonrenewal of a participating teacher. This is not to imply that CTs were eager to recommend nonrenewal or that they did not agonize about such decisions when they had to be made. Nonetheless, CTs rose to the challenge—not in all cases, but at a much higher rate than principals—and when necessary, they recommended nonrenewal. In addition, principals and panel members had confidence in their recommendations, and the teachers' union was part of the process rather than against it. The result was that out of 88 new teachers who were in the program in its first year, 11 (12.5 percent) were not renewed for employment. This included some cases of uncredentialed teachers who were given invitations to return to the district with evidence of a credential and successful teaching elsewhere. In addition, three out of three veterans (100 percent) were encouraged into retirement or into other out-of-classroom responsibilities. In years two through four of the PAR program, the rate of dismissal for beginning teachers fell to 10 percent. Some believed the shift was due to fewer uncredentialed teachers being hired in the first place by the district. In addition, while the veterans placed in the program in its initial year were perceived to be notoriously below standards, by the third year of the program one of the four veterans in PAR that year improved enough to remain in the classroom. This still placed the district below the average of a sample of other established PAR programs, where 30–60 percent of veterans have been remediated.[31]

PAR constituted a major change in accountability when compared with prior dismissal rates in the district. In the year immediately before PAR, only three teachers out of a teaching force of almost 3,000 were not renewed. While some teachers were removed for noninstructional reasons, such as tardiness or drug problems, the union president could not recall (and the district had no record of) any teachers being dismissed for issues of teaching quality in the years immediately prior to PAR.

Far from a draconian or capricious decision, a PAR dismissal represented a concerted and collaborative effort to help a teacher improve that ended with a decision that the teacher's improvement was beyond the ability of the district. Consulting teachers and panel members often noted that they were fulfilling a responsibility to the students of the district, in effect "stepping up" to do a difficult job that had to be done.

Summing Up: A More Professional Model of Teacher Evaluation

The transition to being one's brother's keeper is not easy.[32] The role of consulting teacher is different from that of resource specialist or mentor teacher or other roles that officially elevate teachers into expert status. The gatekeeper function—taking responsibility for decisions about the quality of performance of others in one's profession—is key to being a professional.[33]

The consulting teachers and panel members defined their function as improving the quality of teaching for the clients of the district: students. They expressed a belief that participating teachers could be successful and were committed to helping them get there. If a participating teacher's performance was ultimately not meeting standards, however, they saw their job as recommending dismissal of the teacher. While recommending that someone leave teaching is extremely difficult, consulting teachers mollified themselves with the reminder of the greater good of improving teaching quality for students.

My emphasis on the firing of new teachers as "good news" may seem at best cold-hearted or at worst irresponsible at a time when improving teacher retention is critical to improving teacher quality in urban schools.[34] In a professional model of evaluation that includes a serious concern for client welfare, however, the goal cannot be simply retention. The goal is to retain high-quality teachers (or those who show the potential to grow into high-quality teachers) and to remove from classrooms those teachers who are not performing up to standards and who show little promise of doing so. New teachers are more likely to stay both in teaching and in their current settings if they are provided with the support they need,[35] and the data presented here suggest that PAR may provide that support. New teachers may also take pride in belonging to a profession whose members are seriously engaged in collective responsibility for professional standards.

Lawyers hold collective responsibility for professional standards through the bar. Doctors hold collective responsibility for professional standards through a board. The professional association of teachers, their union, has not historically held any equivalent responsibility. Principals, when asked about important leadership decisions in national surveys, more often reported holding a high level of control over teacher evaluation decisions than over any of the other decisions.[36] An oversight panel for teacher evaluation where more than half the members are teachers and that is co-led by the teachers' union president, however, clearly signals a radical shift in the potential role of teachers and their unions in setting and maintaining standards for the profession. In Rosemont, PAR put the teachers' union and the district, and therefore teachers and administrators, together in a professional community of educators focused on relatively objective measures of the quality of teaching practice. As a result, the school system capitalized on the expertise of teachers in matters of instructional quality, and the teachers' union moved from defending individual teachers to defending the profession of teaching. ☐

Endnotes

1. Donald Boyd, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James H. Wyckoff, "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement" (working paper, Teacher Pathways Project, 2005); and Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73, no. 2 (2005): 417-58.

2. Richard M. Ingersoll, "Four Myths About America's Teacher Quality Problem," in *Developing the Teacher Workforce, 103rd Yearbook of the National Society for the Study of Education*, ed. Mark A. Smylie and Debra Miretzky (Chicago: University of Chicago Press, 2004), 1-33.

3. Camille E. Esch, Christopher M. Chang-Ross, Roneeta Guha, Daniel C. Humphrey, Patrick M. Shields, Juliet D. Tiffany-Morales, Marjorie E. Wechsler, and Katrina R. Woodworth, *The Status of the Teaching Profession 2005* (Santa Cruz, CA: The Center for the Future of Teaching and Learning, 2005).

4. National Commission on Teaching and America's Future, *No Dream Denied: A Pledge to America's Children* (Washington, DC: National Commission on Teaching and America's Future, 2003).

5. Michael A. Copland, "The Myth of the Superprincipal," *Phi Delta Kappan* 82, no. 7 (2001): 528-33; and W. Norton Grubb and Joseph J. Flessa, "'A Job Too Big for One': Multiple Principals and Other Nontraditional Approaches to School Leadership," *Educational Administration Quarterly* 42, no. 4 (2006): 518-50.

6. Linda Kaboolian and Paul Sutherland, "Evaluation of Toledo Public School District Peer Assistance and Review Plan" (unpublished report, John F. Kennedy School of Government, Harvard University, Cambridge, MA, 2005).

7. Madeline Hunter, "Effecting a Reconciliation between Supervision and Evaluation—A Reply to Popham," *Journal of Personnel Evaluation in Education* 1, no. 3 (1988): 275-79.

8. Linda Darling-Hammond, "The Toledo (Ohio) Public School Intern and Intervention Programs," in *Case Studies for Teacher Evaluation: A Study of Effective Practices*, ed. Arthur E. Wise, Linda Darling-Hammond, Harriet Tyson-Bernstein, and Milbrey W. McLaughlin (Santa Monica, CA: RAND Corporation, 1984), 158-66.

9. Karen S. Loup, Joanne S. Garland, Chad D. Ellett, and John K. Rugutt, "Ten Years Later: Findings from a Replication of a Study of Teacher Evaluation Practices in Our 100 Largest Districts," *Journal of Personnel Evaluation in Education* 10, no. 3 (1996): 203-26; and Arthur E. Wise, Linda Darling-Hammond, Harriet Tyson-Bernstein, and Milbrey W. McLaughlin, eds., *Case Studies for Teacher Evaluation: A Study of Effective Practices* (Santa Monica, CA: RAND Corporation, 1984).

10. Yvonne Gold, "Beginning Teacher Support: Attrition, Mentoring, and Induction," in *Handbook of Research on Teacher Education*, ed. John Sikula, 2nd ed. (New York: Macmillan, 1996).

11. For an in-depth discussion of PTs' trust in CTs, see Jennifer Goldstein, "Debunking the Fear of Peer Review: Combining Supervision and Evaluation and Living to Tell About It," *Journal of Personnel Evaluation in Education* 18, no. 4 (2007): 235-52, **http://dx.doi.org/10.1007/s11092-006-9022-3**.

12. Pamela D. Tucker, "Lake Wobegon: Where All Teachers Are Competent (Or, Have We Come to Terms with the Problem of Incompetent Teachers?)," *Journal of Personnel Evaluation in Education* 11, no. 2 (1997): 103-26.

13. Loup et al., "Ten Years Later."

14. Dan C. Lortie, *Schoolteacher: A Sociological Study* (Chicago: University of Chicago Press, 1975).

15. Judith Warren Little, "Assessing the Prospects for Teacher Leadership," in *Building a Professional Culture in Schools*, ed. Ann Lieberman (New York: Teachers College Press, 1988), 78-106.

16. Loup et al., "Ten Years Later."

17. Edwin M. Bridges, *The Incompetent Teacher: The Challenge and the Response* (Philadelphia: Falmer Press, 1986).

18. Dal Lawrence, "California's Opportunity," in *The Peer Assistance and Review Reader*, ed. Gary Bloom and Jennifer Goldstein (Santa Cruz, CA: New Teacher Center, 2000).

19. For an in-depth discussion of the role of the panel, see Jennifer Goldstein, "Designing Transparent Teacher Evaluation: The Role of Oversight Panels for Professional Accountability," *Teachers College Record* 111, no. 4 (2009), **http://www.tcrecord.org/content.asp?contentid=15053**.

20. Richard Elmore, "Education Leadership as the Practice of Improvement" (paper presented at the annual meeting of the University Council for Educational Administration, San Antonio, TX, November 11, 2006).

21. Charles Taylor Kerchner and Douglas E. Mitchell, *The Changing Idea of a Teachers' Union* (New York: Falmer Press, 1988); Charles Taylor Kerchner, Julia E. Koppich, and Joseph G. Weeres, *United Mind Workers: Unions and Teaching in the Knowledge Society* (San Francisco: Jossey-Bass, 1997); Gary Sykes, "Reckoning with the Spectre," *Educational Researcher* 16, no. 6 (1987): 19-21; and Adam Urbanski (presentation at the annual meeting of the Teacher Union Reform Network, Santa Cruz, CA, November 1999).

22. Suzanne R. Painter, "Principals' Perceptions of Barriers to Teacher Dismissal," *Journal of Personnel Evaluation in Education* 14, no. 3 (2000): 253-64.

23. Edwin M. Bridges, *The Incompetent Teacher: Managerial Responses* (Washington, DC: Falmer Press, 1992).

24. Painter, "Principals' Perceptions."

25. Kerchner, Koppich, and Weeres, *United Mind Workers*.

26. Kerchner, Koppich, and Weeres, *United Mind Workers*.

27. Bridges, *Incompetent Teacher: The Challenge and the Response*.

28. Tucker, "Lake Wobegon"; and Loup et al., "Ten Years Later."

29. Tucker, "Lake Wobegon."

30. Tucker, "Lake Wobegon"; and Bridges, *Incompetent Teacher: Managerial Responses*.

31. Darling-Hammond, "Toledo Intern and Intervention Programs"; Philip P. Kelly, "Teacher Unionism and Professionalism: An Institutional Analysis of Peer Review Programs and the Competing Criteria for Legitimacy" (PhD diss., Michigan State University, 1998); Christine E. Murray, "Rochester Teachers Struggle to Take Charge of Their Practice," in *Transforming Teacher Unions*, ed. Bob Peterson and Michael Charney (Milwaukee: Rethinking Schools, 1999); and Denise Hewitt, "The Cincinnati Plan," in *The Peer Assistance and Review Reader*, ed. Gary Bloom and Jennifer Goldstein (Santa Cruz, CA: New Teacher Center, 2000).

32. Little, "Assessing the Prospects"; Patricia A. Wasley, "The Practical Work of Teacher Leaders: Assumptions, Attitudes, and Acrophobia," in *Staff Development for Education in the '90s: New Demands, New Realities, New Perspectives*, ed. Ann Lieberman and Lynne Miller (New York: Teachers College Press, 1991); Charles Taylor Kerchner and Krista D. Caufman, "Lurching toward Professionalism: The Saga of Teacher Unionism," *The Elementary School Journal* 96, no. 1 (1995): 107-122; and Kerchner, Koppich, and Weeres, *United Mind Workers*.

33. Linda Darling-Hammond, "Teacher Professionalism: Why and How?" in *Schools as Collaborative Cultures: Creating the Future Now*, ed. Ann Lieberman (Bristol, PA: Falmer Press, 1990); and John Van Maanen and Stephen R. Barley, "Occupational Communities: Culture and Control in Organizations," *Research in Organizational Behavior* 6 (1984): 287-365.

34. Hamilton Lankford, Susanna Loeb, and James Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis* 24, no. 1 (2002): 37-62.

35. Thomas M. Smith and Richard M. Ingersoll, "What Are the Effects of Induction and Mentoring on Beginning Teacher Turnover?" *American Educational Research Journal* 41, no. 3 (2004): 681-714.

36. Richard M. Ingersoll, *Who Controls Teachers' Work? Power and Accountability in America's Schools* (Cambridge, MA: Harvard University Press, 2003).

Value Added

(Continued from page 27)

a work in progress, a project that is in its adolescence in some respects and its infancy in others. Despite several years of intense work by a number of researchers, we still confront many uncertainties about the statistical and psychometric aspects of value-added models—that is, about the pros and cons of various ways of conducting the analyses and about the limitations of the results. There has been very little research on the practical effects of using VAMs—for example, how teachers' instructional responses compare with those under status or cohort-to-cohort change models. For the time being, using value-added models requires that we choose among alternative approaches with only limited information about the effects that our choices may have on the ratings of teachers or schools, or on the education experienced by students.

Second, we must accept the fact that value-added models, taken by themselves, are not an adequate measure of overall educational quality. Like any other measure based on standardized tests, VAMs provide a valuable but incomplete view of students' knowledge, skills, and dispositions. Because of the need for vertically scaled tests, value-added systems may be even more incomplete than some status or cohort-to-cohort systems. Value-added-based rankings of teachers are highly error-prone. And value-added modeling does nothing to address the interrelated, core problems of an excessive focus on standardized test scores in an accountability system: undue narrowing of instruction, inappropriate test preparation, and the resulting inflation of test scores.

Finally, we have to accept that even within the range of outcomes assessed by the tests used in VAMs, they cannot be counted on to give us true estimates of teachers' value added as opposed to students' overall growth (which has many causes). Although VAMs generally do much better than status and cohort-to-cohort change models in removing the confounding effects of other influences on achievement, we cannot assume at this stage that they will always do this as well as they would have to in order to be trustworthy measures of teachers' effectiveness.

How can we use VAMs in a way that takes these limitations into account and is nonetheless productive? Given the pending reauthorization of NCLB, this is a pressing question. However, given the uncertainties I have described, it should be no surprise that there is no consensus about this. I can only offer my own suggestions:

1. Consider using value-added models rather than cohort-to-cohort or status approaches where appropriate—for example, in elementary school reading and mathematics. But do not let the particular requirements that VAMs impose lead to further narrowing of the accountability system. How much science high school students learn is very important, and if we can't address that with a value-added system, we should address it in some other way.

2. If VAMs will be used, state tests must be constructed from the ground up to be appropriate for this purpose—that is, to support a vertical scale that allows for sensible comparisons from one grade to the next. Efforts to graft VAMs onto grade-specific tests and standards are bad practice.*

3. Use VAMs only with full recognition of the imprecision they entail. Don't pretend that the estimates of teacher or school effectiveness are more precise than they really are. To lessen the impact of this imprecision, add more data, ideally from more years of testing and from other sources entirely. And do not make the consequences of the scores more substantial than the level of precision warrants.

4. Use VAMs primarily to compare classes or schools that start at fairly similar levels of performance. For a number of reasons, comparisons of growth become less and less trustworthy as the initial difference between groups becomes larger. (One reason is that, as explained earlier, the difference between 120 and 140 may not be the same as the difference between 200 and 220.)

5. Don't use test scores as the sole focus of the accountability system. Research in many other fields shows that using too narrow a set of outcomes in an accountability system generates undesirable behavior and distortions in the measured outcome. Evaluations have shown that in the case of test-based accountability systems, these distortions can be severe indeed. VAMs do nothing to lessen this problem.

6. And finally: evaluate, evaluate, and evaluate more. By this, I do not mean testing students more; I mean evaluating the accountability programs themselves. One of the biggest failures of education policy in recent years has been the failure to adequately evaluate the accountability systems that were imposed on teachers and students. We have done enough research to show that the systems do not work as we would like, but we have not done enough to guide the development of better systems. The movement toward VAMs only exacerbates this problem because of the remaining serious gaps in our knowledge of their workings and effects. We need ongoing, independent evaluations to help guide midcourse corrections. For example, we should evaluate the imprecision in value-added estimates, inconsistencies across alternative approaches, the extent of score inflation and other possible biases, and the effects on educational practice and student learning. Our children deserve no less. ☐

* If we want to measure growth well, we will need to put aside standards-based reporting entirely and go back to more traditional scales. This would have other benefits, as the recent change to standards-based reporting was in many respects a bad decision. This is discussed at some length in *Measuring Up*.

For Further Reading

For those interested in reading more about VAM, two sources written for nontechnical audiences are the following:

RAND Corporation. 2004. *The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. Research Brief.* Santa Monica, CA: RAND Corporation. **http://www.rand.org/pubs/research_briefs/RB9050/index1.html**.

Braun, Henry. 2005. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Princeton, NJ: Educational Testing Service. **http://www.ets.org/Media/Research/pdf/PICVAM.pdf**.

A much more detailed but still relatively nontechnical source, which includes discussion of many of the points made here and which was the basis for the RAND research brief noted above, is:

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND Corporation. **http://www.rand.org/publications/MG/MG158**.

AFT Subscription Services

No Gimmicks... Just the Lowest Prices!

AFT+ Member Benefits

Box 258, Greenvale, N.Y. 11548
www.buymags.com/aft

Serving AFT members for over 30 years!

1-800-774-9162

FREE Gift Card on request!

Just For You!

These and Several Hundred others. For Friends and Family too!

1-800-774-9162

www.buymags.com/aft

Publication	Usual Price	Your Price	
Alfred Hitchcock's Mystery	43.90	**29.97**	
Allure	20.00	**11.97**	*
American Baby	23.94	**13.97**	
American Heritage	24.00	**24.00**	
American History	25.00	**25.00**	
American Photo	21.00	**10.50**	
Aquarium Fish	27.97	**14.99**	
Architectural Digest	42.00	**24.00**	*
Arthritis Today	30.00	**12.95**	
Artist's Magazine	19.96	**19.96**	
Asimov's Science Fiction	43.90	**32.97**	
The Atlantic	29.95	**24.95**	
Backpacker	29.98	**19.94**	*
Baseball Digest	35.94	**19.95**	
Better Homes & Gardens 1 yr	22.00	**14.97**	
2 yrs		**22.00**	
Bicycling	19.98	**19.94**	*
Bird Talk	27.97	**13.99**	
Black Enterprise	22.00	**17.95**	
Boating	28.00	**14.00**	
Bon Appetit	28.00	**24.00**	*
Brides	18.00	**11.97**	
BusinessWeek [50 iss]	58.79	**40.00**	*
Cabella's Outfitter Journal	12.99	**9.99**	

3 Special Rates for our members!

Enjoy them all year long!

Car & Driver	22.00	**10.97**	
Cat Fancy	27.97	**14.99**	
Coach, Scholastic	23.95	**17.95**	
Coastal Living	36.00	**16.00**	
Columbia Journalism Review	20.00	**11.95**	
Computer Shopper	25.00	**14.99**	
Conde Nast Traveler	19.97	**14.97**	*
Consumer Reports	29.00	**29.00**	
Cooking Light	18.00	**18.00**	
Cosmopolitan	29.97	**18.00**	*
Country Living	24.00	**12.00**	
Cruise Travel	35.94	**17.95**	
Dell Original Sudoku	20.96	**15.97**	
Discover	29.95	**24.95**	
Disney and Me (ages 2-6)	29.70	**24.97**	
Disney's Princess (age 4+)	39.60	**29.97**	
Dog Fancy	27.97	**14.99**	

* These rates for AFT members and college students only.

Ebony	20.00	**14.97**	
Economist	125.00	**77.00**	*
Elle	28.00	**14.00**	
Elle Decor	29.00	**14.50**	
Ellery Queen's Mystery	43.90	**29.97**	
Entertainment Weekly [57 iss]	39.95	**34.95**	*
ESPN	26.00	**13.00**	
Esquire	15.94	**8.97**	
Essence	22.00	**18.96**	

The latest ideas, events & current issues

A best buy $29.95

Every Day with Rachael Ray	34.00	**24.00**	
Family Circle	19.98	**11.97**	
Family Fun	16.95	**9.95**	
Family Handyman	24.00	**15.00**	
Field & Stream	19.97	**10.00**	
First For Women	19.97	**19.97**	
Fitness	19.98	**12.00**	
Food & Wine	32.00	**19.00**	
Forbes	59.95	**24.95**	
Foreign Affairs	44.00	**32.00**	
Foreign Policy	24.95	**17.50**	
Fortune [25 iss]	59.95	**29.98**	*
Girls Life	19.95	**14.95**	
Glamour	20.00	**11.97**	*
Golf	19.95	**15.00**	
Golf Digest	27.94	**17.77**	
Golf World	53.97	**29.97**	
Golfweek	79.95	**30.00**	
Good Housekeeping 1 yr	23.97	**12.00**	
2 yrs		**24.00**	
Gourmet	20.00	**15.00**	
GQ	20.00	**15.00**	*
Hallmark	18.95	**14.95**	
Harper's Bazaar	18.00	**15.00**	
Harper's Magazine	21.00	**11.97**	
Hawaii	20.00	**15.00**	
Hispanic Magazine	24.00	**10.00**	
Home	24.00	**12.00**	
House Beautiful	19.97	**12.00**	
Humpty Dumpty (ages 4-6)	22.95	**17.29**	
Inc. (18 issues)	19.00	**14.00**	
Instructor (K-8)	19.95	**9.95**	
InStyle [12 issues]	22.15	**18.00**	*
Interview	14.97	**9.97**	
Jet	38.00	**24.00**	
The Kiplinger Letter	79.00	**48.00**	
Kiplinger's Personal Finance	23.95	**12.00**	
Kiplinger's Retirement Report	59.95	**29.95**	
Ladies Home Journal	16.97	**12.00**	

Visit our website at

Latina	17.97	**9.97**	
Lucky	20.00	**14.97**	
Marie Claire	19.97	**12.00**	
Martha Stewart Living	28.00	**24.00**	
Men's Fitness	21.97	**10.00**	
Men's Health	24.94	**24.94**	*
Men's Journal	19.97	**9.97**	
Men's Vogue	18.00	**15.00**	

NOW! 53 weekly issues at a special member rate.

53 issues $20.00 106 issues $39.97

Metropolitan Home	29.00	**13.97**	
Midwest Living	19.97	**11.97**	
Modern Bride	17.97	**9.95**	
Money [12 issues]	36.82	**19.95**	*
More Magazine	20.00	**14.97**	
Mother Earth News	18.00	**13.97**	
Motor Trend	18.00	**10.00**	
Motorboating	19.97	**14.00**	
The Nation	52.00	**26.00**	
Natural Health	23.90	**11.97**	
Natural History	30.00	**25.00**	
New York 1 yr	29.90	**14.97**	
2 yrs		**29.90**	
New Yorker	52.00	**29.95**	*
Newsweek [53 iss]	41.87	**20.00**	*
[106 iss]		**39.97**	*
O Oprah 1 yr	28.00	**18.00**	*
2 yr		**36.00**	
Old House Journal	27.00	**13.97**	
Organic Gardening	24.96	**24.96**	*
Outdoor Photographer	19.94	**10.98**	
Outside	19.95	**16.00**	
Parenting	17.97	**9.97**	
Parents	15.98	**9.97**	

Unbeatable LOW rate on Economist a member favorite!

Full year - JUST $77.00

Publication	Usual Price	Your Price	
Self	20.00	**14.97**	*
Seventeen	19.95	**12.00**	
Shape	21.97	**14.97**	
Ski [7 Iss] or Skiing [6 iss]	14.97	**11.00**	
Smart Money	24.00	**15.00**	
Smithsonian [12 iss]	34.00	**12.00**	
Sound & Vision	29.00	**12.00**	
Southern Accents	26.00	**18.00**	
Southern Living	36.00	**19.95**	
Sparkle World (girls 3+)	39.60	**29.97**	
Sports Illustrated	89.04	**39.95**	*
The Weekly Standard	79.96	**47.96**	
Stereophile	19.94	**10.00**	
Sunset	24.00	**19.95**	
Teen Vogue [10 iss]	15.00	**10.00**	
Tennis	18.00	**12.00**	
This Old House	20.00	**20.00**	
Time [56 issues]	59.95	**29.95**	*
Time Out - New York	39.94	**19.97**	
Town & Country	28.00	**15.00**	

Oprah, now available at a low, low rate. Just for our members!

12 issues $18.00

- *Best Titles*
- *LOWEST Rates*
- *GREAT GIFTS!*

Extended Office Hours
Mon.-Thur. 9am-7pm
& Fri. til 5pm ET

PC Magazine	44.97	**24.97**	
Popular Mechanics	24.00	**24.00**	
Popular Photography	24.00	**11.97**	
Popular Science	19.95	**15.94**	
Portfolio	20.00	**12.00**	
Preschool Playroom	29.70	**24.97**	
Prevention	21.97	**16.94**	*
Psychology Today	21.00	**15.97**	
Reader's Digest	24.98	**13.96**	
large print edition	29.96	**21.95**	
Selecciones (*español*)	27.46	**19.97**	
Real Simple [12 iss]	28.95	**24.00**	*
Redbook	17.97	**12.00**	
Road & Track	22.00	**10.97**	
Rolling Stone	25.94	**14.95**	
Runner's World	24.00	**20.00**	*
Salt Water Sportsman	24.97	**20.00**	
Scientific American	34.97	**24.97**	
Scuba Diving	20.15	**11.97**	
Travel & Leisure	45.00	**19.00**	
TV Guide	58.14	**39.96**	
U.S. News [1 yr]	49.75	**20.00**	
US Magazine	65.00	**52.00**	
Vanity Fair	28.00	**24.00**	*
Veranda	24.00	**18.00**	
Vogue	29.95	**17.97**	
W Magazine	29.90	**14.95**	
Weight Watchers	17.70	**13.95**	
WildBird	19.97	**12.99**	
Wine Enthusiast	35.00	**26.95**	
Wired	24.00	**12.00**	
Woman's Day	18.00	**9.99**	
Women's Health	19.95	**14.97**	*
Working Mother	10.38	**9.97**	
Writer's Digest	19.96	**18.96**	
Yachting	19.97	**16.00**	
Yoga Journal	21.95	**15.95**	

www.buymags.com/aft

Hundreds of Others Just Ask!

For renewals include a mailing label, if available. All subscriptions one year unless otherwise indicated

AFT SUBSCRIPTION SERVICES
Box 258 • Greenvale, NY 11548

Name_____

Address_____

City, State, Zip_____

email _____

Your School_____

Home Phone (_____)_____

FREE gift card upon request-- please send us a separate note.

Publication Name	Years	Price
	Total	

❑ Check enclosed payable to: **AFTSS**
❑ Charge to my credit card
❑ Visa ❑ MasterCard ❑ Discover ❑ Amex

Acct: _____ Exp. Date: ____

❑ Please bill me (phone # required)

S2809

For students to learn to their potential, schools must be safe and orderly. But preventing behavior problems—and handling them effectively—are easier said than done. For effective strategies, visit the Tools for Teachers section of the AFT's Web site at **www.aft.org/tools4teachers/defining-consequences.htm**. And for a helpful reminder of appropriate behavior throughout your school, request a copy of the poster shown here—the first one is free; additional copies are 50 cents each. Contact the AFT Order Department, 555 New Jersey Ave. N.W., Washington, DC 20001. If ordering multiple copies, include a check payable to the American Federation of Teachers.