Value-Added Tests: Buyer, Be Aware

The value-added assessment model is one over-the-counter product that may be detrimental to your health.

Audrey Amrein-Beardsley

Who doesn't laugh when a drug commercial presents a clip of a young, otherwise happy and healthy person laughing and flying a kite at the park and then dramatically exposes some ailment—only to fix it by unveiling a prescription drug, along with its potential side effects? What is laughable is that the side effects often seem worse than the problem itself.

I may have found more humor in these commercials than others have because I grew up in a family opposed to even over-the-counter drugs. Drugs were simply not a part of my family's holistic approach to healthy living—until last year, when I discovered I had a heart condition. I was prescribed a drug cocktail consisting of six medications, three of which carry serious side effects. No longer was I laughing at the poor souls portrayed in those drug commercials. Thanks to the Food and Drug Administration (FDA), I quickly became an educated consumer.

The FDA is the oldest and most respected protector of wellness in the United States. It exists to guarantee that no harm is done to consumers of foods and drugs. Specifically, it ensures that the benefits of the foods and drugs it approves outweigh the risks they pose and that their benefits and risks, once scientifically documented, are fully disclosed to the public to enable consumers to make wise health decisions.

Might the FDA approach also serve as a model to protect the *intellectual* health of the United States? Might this be a model that legislators and education leaders follow when they pass legislation or policies whose benefits and risks are unknown? Don't students, teachers, and administrators in U.S. public schools deserve similar protection? In light of these questions, let's look at one suggested "cure" for what's ailing our schools.

Take the Model—and Call Me in the Morning

Currently, NCLB mandates that all U.S. states measure student learning using standardized achievement tests. This is not likely to change. NCLB also requires states to report on school progress using adequate yearly progress (AYP) measures. But because of AYP's shortcomings, some states now receive funds to integrate value-added assessment models into their accountability procedures, largely to help states comply with the accountability provisions written into NCLB.

Value-added models assess teachers, schools, and districts on the value they add to student learning, from the moment students enter the classroom to the time they leave. In theory, this makes more sense than just capturing where students are academically at the end of each school year.

But it is far from certain that value-added models work in the ways theorized. There is a risk associated with

Might the FDA approach also serve as a model to protect the *intellectual* health of the United States?

blindly adopting the value-added models—they may be detrimental to consumer health. Just as the FDA regulates foods and drugs, we need a Federal Education Agency to provide the science-based, accurate information that educators need to be informed consumers. Such an agency might warn consumers about the benefits and risks of the value-added assessment models currently "prescribed." To protect the public good, such an organization might examine whether the most popular, widely adopted, sophisticated, and expensive "over-the-counter" value-added model—the Education Value-Added Assessment System (EVAAS) developed by William L. Sanders—really measures up. The model has three limitations.

Limitation 1: A Reliance on Standardized Tests

The EVAAS model relies on standardized tests to measure levels of change in student learning. It's unclear whether standardized tests can accurately measure what students know and are able to do at one point in time, let alone over time to measure "knowledge added." The effect of districts, schools, and teachers on student learning is also unclear; we need to consider whether

it's possible to attribute gains or losses in student achievement solely to the quality of instruction, independent of other school and life factors. This is the fundamental assumption on which the system relies, which, if we were to seriously consider it, might cause a model recall.

Test Data Irregularities

The EVAAS model requires complete and high-quality longitudinal test data that most states currently do not have. Students sometimes miss tests. Student test score data are often not linked to teacher names. Students are often misreported by class and grade level. And sometimes data show that students jump from the top to the bottom of the class, or vice versa, from one year to the next, which is nearly impossible in actuality.

Data errors like these, which are often caused by student mobility, missing test scores, data-processing errors, or students incorrectly bubbling in their score sheets, affect thousands of student records. Model developers claim these things don't matter, that the system can operate regardless.

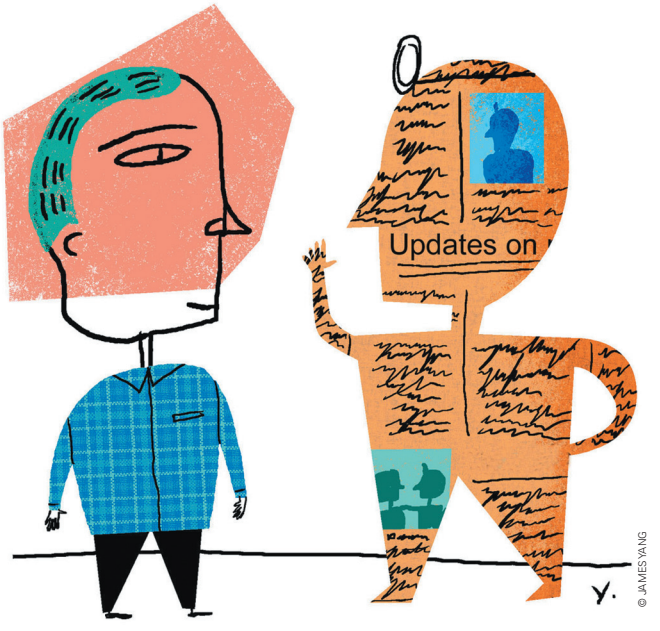
Student Risk Factors

The EVAAS model does not control for student risk factors, making it the only sophisticated assessment model that does *not* account for such things as family income, ethnicity, and other student background variables.

Developers of the model state that the effects of these factors on student growth are negligible. Yet educators know too well that student background variables unquestionably affect student achievement and the progress students make from year to year. How could the achievement gap continue to widen if these factors play no role?

Class Size

Statistical errors in test results frequently occur when fewer students are in a given class, a problem that prevents truthful claims about the quality of teachers with class sizes below a certain number. In the EVAAS model, general and special education teachers who teach smaller classes are more likely assumed to be average. An ineffective teacher who teaches a large class might be penalized for being below average, whereas an equally ineffective teacher who teaches a smaller class may go undetected. The larger the class, the more "accurate" the estimate. This model makes the process of evaluating teacher quality unfair, discriminating against teachers who have larger classes.



Grades and Subjects Tested

Only students in certain grade levels must take the standardized accountability tests. This situation subjects teachers to accountability measures in some grades, but not in others. In addition, many of these tests only assess students' reading and mathematics skills, exempting teachers who teach other subjects from being held accountable in similar ways.

Teacher Effect

The EVAAS model is also incapable of controlling for out-of-school learning and the effects one teacher might have on another. Let's say students complete a standardized test in the spring in one teacher's classroom. They complete the school year still learning from that teacher, spend three months in the summer losing or gaining variable amounts of knowledge, enter the classroom of a new teacher in the fall, and then take the "post-test" the following spring under the tutelage of the new teacher. It is impossible to prove that the losses or gains posted from the previous year to the next are solely a result of the current teacher's efforts. Although system developers argue that their system can factor these effects out, this remains unclear—and unlikely.

The issue becomes more convoluted when students enter middle and high school and switch teachers and classrooms daily, sometimes taking classes in the same subject areas during the same semester. For example, if a student is taking geometry and algebra the same semester, who is to say that the geometry teacher was more or less effective than the algebra teacher or that the value the geometry teacher added to the student's learning about math had nothing to do with what the student learned in algebra? Who is to say that a

student who switched language arts teachers midsemester learned more about reading from one teacher than the other? What about teachers who team teach or teach in other atypical classroom settings? The model neglects this complexity.

Student Assignment

And what does all this mean for evaluating teacher effectiveness when students are not randomly placed into classes? If one high-quality teacher gets an amazing set of students and another equally effective teacher gets an unexceptional set, the students of the first teacher will most likely learn more within one year, and their teacher will be unfairly rewarded as a higher-quality teacher. This situation is more likely in schools in which assertive parents push their children into "better" classrooms. Conversely, if a teacher is assigned a disproportionate amount of difficult-to-teach students—possibly because the principal believes that he or she can teach at-risk students more effectively—and these students gain less than other students in comparable classrooms, is it fair to say the teacher is less capable, successful, or qualified than other teachers?

We can say that one teacher caused students to learn more than another one did only if we randomly assign students into classrooms. The same holds true for statements about schools and districts. Most value-added researchers agree with this, yet some continue to use data from their models to make consequential decisions about teachers, schools, and districts.

A Single Indicator

At best, the EVAAS model might be useful at face value to help identify teachers who need professional development or schools and districts in need of intervention if, *and only if*, value-added score reports are not used in isolation from other data confirming that the teachers, schools, or districts are, in fact, struggling to succeed (see also Bracey, 2007).

The use of one single indicator to make consequential decisions about students, teachers, schools, or districts violates the first of the 12 Standards for Educational and Psychological Testing set forth by the American Educational Research Association (AERA), the American Psychological Association, and the National Council on Measurement in Education (AERA, 2000). These standards represent the professional consensus on the appropriate uses of tests.

We can say that one teacher caused students to learn more than another one did only if we randomly assign students into classrooms.

Limitation 2: Lack of Evidence of Validity

The model's developers state that adopting the EVAAS model will make visible certain education findings that were indiscernible in the past (Sanders & Horn, 1994) in fair, objective, and unbiased ways (SAS, 2007). Without these findings, they claim, the real effects on student learning would continue to go unaddressed (Sanders, 1998). Purportedly, the model will help districts and schools make data-informed decisions that will ultimately increase student performance.

Also, proponents state that "combining value-added analysis and improved high school assessments will lead to improved high school graduation rates, increased rigor in academic content, higher college-going rates, less college remediation, and increased teacher accountability" (Battelle for Kids, n.d.). But nowhere do the developers provide evidence to substantiate these claims.

The model's developers have used their value-added data to notify parents of the chances their children will or will not pass upcoming tests or graduate from high school. They have also used the system to predict students' scores on college entrance exams, estimate the likelihood students will get into state colleges and universities, predict which students are more suited to technical majors, and determine the probability of students receiving *A*s and *B*s their freshman year in college.

Using inexact data to predict things about students' lives is unethical, unprofessional, and borders on education malpractice. Many parents and teachers already think they know which students are at risk; let us not rely on imperfect statistics to notify high-achieving students that they are free and clear or remind low-achieving students that the odds are against them. Making such predictions may directly or indirectly cause them to come true.

The model's developers also claim that because their product singles out teachers whose students post either above- or below-average gains, it's the best tool out there for rewarding or penalizing teachers. Yet the developers have conducted no studies to examine

whether teachers determined as highly effective are also (1) teachers with more years of experience, (2) teachers whose supervisors or peers would also be classified as highly effective, (3) teachers who received high scores on their teacher licensure tests, (4) teachers who have higher levels of education, or (5) teachers who have received teaching awards and honors, are National Board certified, and the like.

Moreover, personnel in the districts and schools that have implemented the model do not seem to be using the data in the expected and promoted ways. This is largely because of the confusing data reports and a lack of professional development opportunities to help teachers and administrators understand the model's output.

Limitation 3: Lack of Transparency

There has been insufficient external examination of the EVAAS model to inform recommendations or regulatory decisions about its use, benefits, and risks. The question here is whether there have been enough empirical studies conducted to warrant the federal and state education policies mandating the use of this system.

The model's developers have not completely opened up their system—in particular, the computational algorithms used to analyze test data—to external or peer review. Nor have they released any value-added data they have collected to enable other researchers to verify the claims they make. This makes scientific research by external statisticians nearly impossible, limiting researchers' capacity to make sound recommendations about the model to inform education policies and provide consumers with the facts they need to make their own "regulatory" decisions.

In 1997, developers asserted that they had undertaken "extensive efforts" to increase understanding of the system, formerly known as the Tennessee Value-Added Assessment System (TVAAS), and they explained the system in great detail. They also stated that "detailed external reviews from both the statistical and educational evaluation communities have confirmed that the properties of the TVAAS results are as claimed" (Sanders, 1998, p. 26)—but they didn't provide citations or references to these external reviews. Four sets of external reviewers examined the assessment system in depth: Two reviewers praised the system, one reviewer raised significant points of contention, and the last reviewer was one of the model's developers (Sanders & Wright, 2008).

This model discriminates against teachers who have larger classes.

Educating the Education Consumer

In all fairness, all value-added models are flawed, especially when it comes to their reliance on standardized tests and the assumptions about what these tests can reveal. The EVAAS model is the most sophisticated, or the least inferior, of these models.

Nevertheless, should the issues that contaminate the practicality of the EVAAS model warrant its removal from the market? Yes, at least until external reviewers can verify the model's assumptions about what standardized tests can reveal, validate the inferences drawn about students and teachers, begin necessary internal and external research studies, answer commonsense questions, and inform consumers about the system's benefits and risks.

We need to take our education health as seriously as we take our physical health. Education consumers should get to know the model before education policymakers force them to blindly accept it, simply because the theory behind it makes sense. And they should have the opportunity to learn about the benefits and risks of the EVAAS approach because, in the end, they—and not the software developers or the system builders—will experience the side effects. EL

References

American Educational Research Association (AERA). (2000). *AERA position statement on high-stakes testing in PreK-12 education.* Available: www.aera.net/?id=378

Battelle for Kids. (n.d.). *High-school value-added project.* Retrieved February 1, 2007, from www.battelleforkids.org.

Bracey, G. W. (2007, May 1). Value subtracted: A "debate" with William Sanders. *The Huffington Post*. Available: www.huffingtonpost.com/gerald-bracey/value-subtracted-a-debate_b_47404.html

SAS. (2007). *Dr. William L. Sanders*. Available: www.sas.com/govedu/edu/bio_sanders.html

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11), 24–27.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299–311.

Sanders, W. L., & Wright, S. P. (2008, April 14). *A response to Amrein-Beardsley (2008): "Methodological concerns about the Education Value-Added Assessment System."* Available: www.sas.com/govedu/edu/services/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf

Audrey Amrein-Beardsley is Assistant Professor in the College of Teacher Education and Leadership at Arizona State University, Phoenix; audrey.beardsley@asu.edu.