THE LIMITED UTILITY OF VALUE-ADDED MODELING IN EDUCATION RESEARCH AND POLICY

JIMMY SCHERRER, NORTH CAROLINA STATE UNIVERSITY

In recent years, there has been a push to create new teacher accountability systems. A large continuum of designs has been proposed. At one end, designers claim that we need to identify ineffective teachers and remove them from the profession. At the other end, designers suggest that we spend our energy identifying effective teachers and reward them (with bonuses or tenure). Despite the varying foci, each design faces the same challenging question: *How do we measure effectiveness*? It is becoming increasingly common to use the statistical technique *value-added modeling* to help answer this question. However, it appears that the utility of this metric is still not fully understood by those who are adopting it.

This is troublesome—if you adopt the wrong metrics, you will strive for the wrong things. The field of education is in a precarious situation: the use of value-added modeling in accountability systems is outpacing the understanding of how to correctly process the estimates that it generates.

I. VALUE-ADDED MODELING

In theory, value-added modeling (VAM) captures a fraction of achievement growth over time that can be attributed to a particular teacher. The models achieve this by controlling for non-school factors related to student achievement (e.g., socioeconomic status, health, parent education). These "controls" allow for a computation of *expected growth* for each student. Any deviation from this expected growth is attributed to the teacher. Averaging these deviations across many years for many students, a teacher is given a value-added estimate.

This technique is attractive to education researchers and policy makers because, in theory, the estimates produced can be used as counterfactual quantities—that is, one can estimate how students of Teacher A would have done if they had been taught by Teacher B. Howbeit, it should be noted, these counterfactual quantities rest on the underlying assumptions of the statistical models. Recent research has illustrated that the underlying assumptions of VAM are violated in current educational settings. For example, one underlying assumption of VAM is that school administration and collaboration among teachers do not have a significant impact on student achievement. However, in reality, we know that a teacher's surrounding professional community inevitably affects the kind of teaching that students are offered.[1]

Even if one argues that the violation of the underlying assumptions of VAM are not that egregious, as many accountability enthusiasts do, there are further issues of using this technique in education accountability systems that must be brought into current discourse. In this article I highlight two of them: the reliability of the models' estimates and the validity of the inferences made based on the estimates.[2]

II. THE RELIABILITY OF THE ESTIMATES

The reliability of a measure refers to the stability of the estimates that are generated. With respect to value-added modeling in teacher accountability systems, achieving stability would mean that the models consistently estimate the same teachers to be effective and the same teachers to be ineffective. A close examination of commonly used value-added models reveals that this is not the case. For example, a study by Koedel and Betts (2005) found significant fractions of teachers moved up or down by two quintiles or more when the sample of years was shifted. Only 35% of teachers in the top quintile during one academic year remained in the top quintile the following academic year. That is, the estimates generated by the models implied that the performance of 65% of a school's most "effective" teachers decreases from one year to the next. This non-persistence of performance implies non-persistent sources of error and raises grave reliability issues.

Figure 1 helps conceptualize why value-added estimates are so unstable. The figure illustrates one teacher's value-added estimates (represented by the red dots) across two years. Although underpublicized, these estimates include considerable margins of error. This error is represented by the green line in Figure 1. The correct way to interpret this teacher's value-added estimate for year 1 is as follow: The model estimates that this teacher is at the 60th percentile, but she could be as high as the 70th percentile or as low as the 50th percentile. For year 2: The model estimates that this teacher is at the 70th percentile, but she could be as high as the 80th percentile or as low as the 60th percentile.

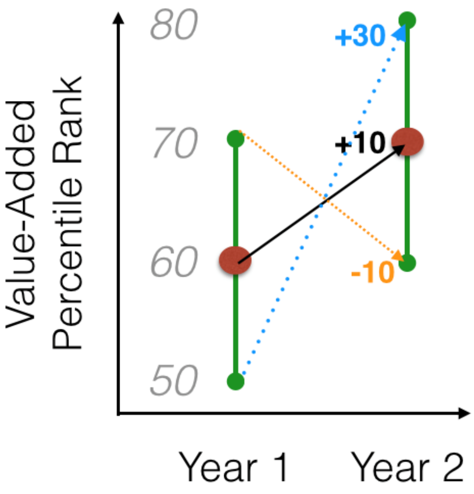


Figure 1 Value-added percentile rank for one teacher across two years. The estimate is represented by the red dot. The estimate's error is represented by the green line.

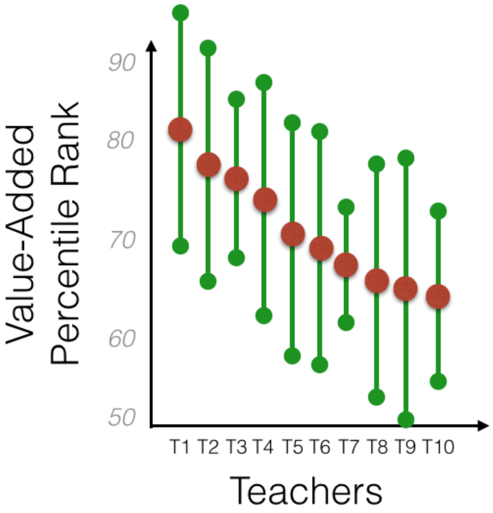
If the estimates are correct, this teacher's percentile rank increased 10 percentile points from year 1 to year 2 (represented by the solid black line). But she may actually have increased by 30 percentile points (represented by the dotted blue line). Conversely, she may have decreased by 10 percentile points (represented by the dotted orange line). When the interpretation of value-added estimates give attention to the estimates' error, it is easier to understand why the estimates are so unstable, and thus unreliable.[3]

Similarly, this error helps conceptualize why VAM cannot be used to rank teachers by their "effectiveness." Figure 2 includes the value-added estimates of 10 teachers in a given school. When only the estimate is considered, as is often the case in current accountability discussions, it seems plausible to rank teachers. But when attention is given to the margins of error, it becomes clear that the estimates generated through VAM are not precise enough to rank teachers. In Figure 2, for example, when the error bar is taken into consideration, the 10th teacher listed might actually be more "effective" than the 1st teacher listed.

III. THE VALIDITY OF THE INFERENCES

Using the estimates generated by value-added models as a proxy for teacher effectiveness is dodgy for multiple reasons. First, there is a lot more to effective teaching than getting students to master the content that appears on standardized assessments (e.g., making connections between the content being explored and its real-world application, creating a

Figure 2 Value-added estimates and error bars for 10 teachers. The red dots represent the actual value-added estimates. The green lines represent the estimates' errors.



climate where students feel comfortable expressing their opinion, and so on). Further, most standardized assessments only test mathematics and literacy. As a result, to ensure high marks on the assessment, the curriculum is being narrowed as teachers increase the amount of time that they teach these two domains. This increased dosage of mathematics and literacy is at the expense of other domains, such as science and history.

There is also narrowing *within* the mathematics and literacy curricula. Standardized tests primarily focus on basic skills and knowledge and not the high-level conceptual understanding and reasoning that are emphasized in contemporary standards documents. As a result, teachers are increasing the amount of time that they drill their students with the basic facts and knowledge that appear on the test (often at the explicit direction of their administrators). This hyper focus on what appears on the test is at the expense of other important (non-tested) skills such as sense-making and reasoning.

The narrowing *of the curriculum* and the narrowing *within the curriculum* should not come as a surprise. The higher the stakes are for improved performance, the more likely behavior will change in unintended ways. In the social sciences, we refer to this phenomenon as Campbell's Law. For those interested in putting a lot of weight behind value-added estimates in education accountability systems, examples of Campbell's Law at work in other sectors are instructive.

Health Care

During the late 1980s in the United States, New York City surgeons and their affiliated hospitals began receiving public "report cards" based on the mortality rate of their patients. One particularly interesting case involved a hospital that was put on probation for receiving a low "grade" one year, but received the highest "grade" in the subsequent year. To accountability enthusiasts, the public shaming that accompanied the low "grade" seemed to work: surgeons must have changed their behavior in a manner that resulted in improved surgeries. It turns out that surgeons did indeed change their behavior, but in unintended ways.

Further investigation found that, instead of improving their surgery practices, surgeons simply began to turn away the sickest patients—a patient that dies at home, or on someone else's table, does not "count" against them on their report card. This is not an isolated instance. A 2005 New York Times article reported that 79% of surgeons (who responded to their survey) admitted that the fear of receiving a low ranking affected their decision about whether to perform a surgery.

Law Enforcement

The effectiveness of police departments is measured by, among others, a department's clearance rate—the percent of crimes that have been solved. In a well-known, and widespread, tactic, police departments offer interesting plea deals to criminals when a department's clearance rate is too low. The tactic involves reduced sentences for criminals who "confess" to crimes that they did *not* commit (which efficiently "solves" unsolved crimes). Education is not immune to this type of behavior. The current narrowing of the curriculum and the narrowing within the curriculum, as well as recent major cheating scandals, should come as no surprise. When there are high stakes attached to increasing a quantitative measure, the measure usually does indeed increase, but often in very distorted ways.[4]

The other issue related to validity is the increasing practice of using value-added estimates as a proxy for effective teaching. Different teaching practices reflect different pedagogical epistemologies. These epistemologies are rooted in various learning perspectives. Table 1 lists three common perspectives on learning: the behaviorist, the cognitivist, and the situative. Each perspective is associated with a different grain size. The behaviorist perspective focuses on basic skills, such as arithmetic.

The cognitivist perspective focuses on conceptual understanding, such as making connections between addition and multiplication. The situative perspective focuses on practices, such as the ability to make and test conjectures. Effective teaching includes providing opportunities for students to strengthen each focus. However, traditional standardized assessments mainly contain questions that are crafted from a behaviorist perspective. The conceptual understanding that is highlighted in the cognitivist perspective and the participation in practices that is highlighted in the situative perspective are not captured on traditional standardized assessments. Thus, the only valid inference that can be made from a value-added estimate is about a teacher's ability to teach the basic skills and knowledge associated with the behaviorist perspective.

The columns in Table 1 represent the three different learning perspectives. The rows describe the learning, teaching, and assessing that are associated with each perspective. When using assessment data to make an inference about classroom teaching, there needs to be coherence and consistency within a column. The current practice of using value-added estimates as a proxy for effective teaching introduces a "leap" across perspectives. That is, scores from traditional standardized assessments rooted in behaviorism are being used to make inferences about classroom teaching practices that are coherent with different perspectives. This "leap" essentially eliminates the ability to make a connection between high value-added estimates and current notions of effective classroom teaching.

"Effective teaching includes providing opportunities for students to strengthen each focus. However, traditional standardized assessments mainly contain questions that are crafted from a behaviorist perspective."

Table 1 Three perspectives of learning and their connections to teaching and assessment

	BEHAVIORIST	COGNITIVIST	SITUATIVE
LEARNING	Accumulating and strengthening associations and skills	Increased meaning making	Increased ability to participate in the practices of a particular community
TEACHING	Direct instruction of associations and skills	Help students build relationships between new knowledge and old knowledge	Negotiate activity settings that provide students with opportunities to participate in various practices
ASSESSMENT	Application of associations and skills: Assessments of independent samples of knowledge to estimate how much of a domain a student has acquired	Questions/tasks about whether students understand general principles in a domain: Questions that ask students to explain or justify, or multi-day tasks that require students to apply multiple skills to solve a problem	Observation of practices by individuals and groups (often *in situ*)

IV. RECOGNIZING THE LIMITATIONS

The use of value-added modeling by educational researchers and policymakers can be quite useful. For educational researchers, the technique can help identify overarching trends and help isolate the effect of new interventions. For educational policymakers, accountability systems that utilize value-added modeling are a significant improvement to traditional systems that utilize status measures. Status measures judge teachers based on their students' raw assessment scores (without controlling for factors known to have an effect on outcomes), which rewards teachers for *who* they teach rather than *how* they teach.

Despite these attractive qualities, the utility of value-added modeling is limited. As discussed in this paper, the technique is simply not precise enough to rank teacher performance. Moreover, it is recognized that teachers can improve their value-added estimate without actually improving their classroom teaching (i.e., Campbell's Law). Thus, in robust accountability systems, serious attention needs to be given to how much weight a value-added estimate will have. In my opinion, the recent trend of using a teacher's value-added estimate as 50% of his or her evaluation is much too high.

JIMMY SCHERRER, PhD, is an assistant professor at North Carolina State University, College of Education. His research focuses on how different learning environments enable or constrain opportunities to learn and identify, how educational policy decisions shape learning environments, and how social interactions within environments mediates learning.