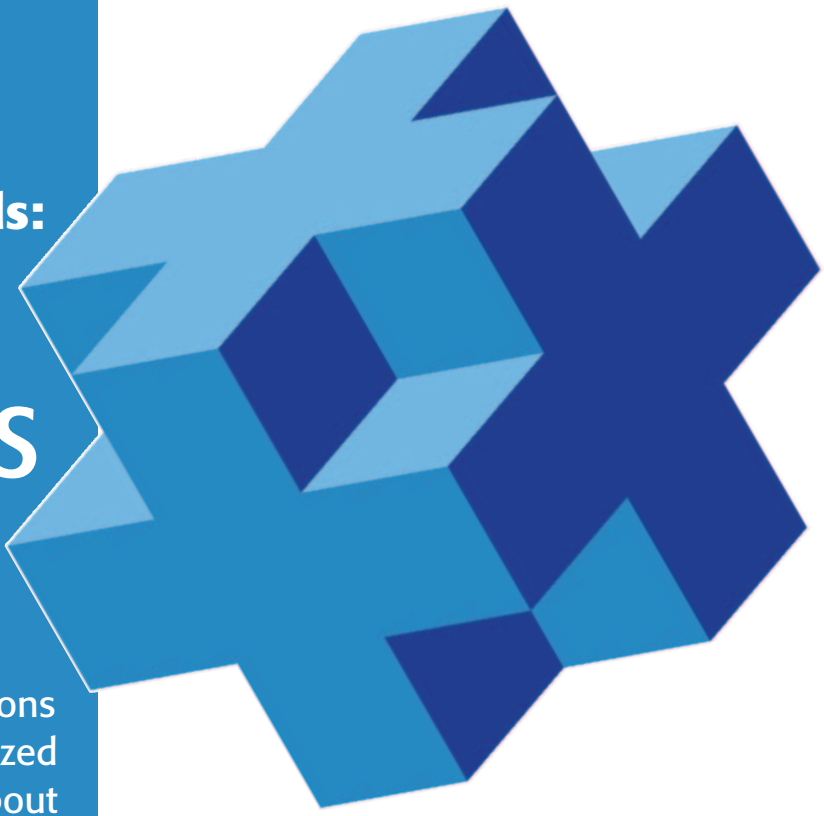
Value-added models:

What the experts say

Researchers share the opinions of scholars who have analyzed and written extensively about value-added models.



By Audrey Amrein-Beardsley, Margarita Pivovarova, and Tray J. Geiger

When a collection of experts, people with special knowledge and skills pertaining to a problem, are all skeptical about a proposed solution to that problem, it's a good idea to hear what they say — and, perhaps, find a better solution.

The problem in question? Value-added models (VAMs). For the past decade, expert statisticians and econometricians have been exploring alternative methods to this approach to documenting teacher performance. Its use is now widespread. Before the passage of the Every Student Succeeds Act (ESSA) in 2015, 44 states and the District of Columbia had implemented high-stakes policies to evaluate teacher effectiveness based on VAMs. Although ESSA has since curbed the extent to which states are adopting and implementing VAMs (as in Alabama, Georgia, and Louisiana), in other states (such as New Mexico, New York, and Texas), VAMs are still playing substantial roles in their teacher-level accountability efforts.

AUDREY AMREIN-BEARDSLEY (audrey.beardsley@asu.edu) is a professor in the Mary Lou Fulton Teachers College, Arizona State University, Tempe, Ariz., where **MARGARITA PIVOVAROVA** (Margarita.Pivovarova@asu.edu) is an assistant professor, and **TRAY J. GEIGER** (tjgeiger@asu.edu) is a doctoral candidate.

VAMs should identify outlier teachers, schools, and classrooms, which educators could study as positive (or negative) examples.

Practice versus theory

Value-added models are designed to measure how much value a teacher purportedly adds to (or detracts from) students' growth as evidenced on large-scale standardized achievement tests over each school year. The models statistically control for students' prior testing histories, with some also controlling for student-level variables (for example, demographics, English language status, or special education status) and school-level variables (such as class size or school demographics). In theory, measuring teachers' value-added allows for richer analyses of standardized test score data because groups of students are followed to assess their learning trajectories from the time they enter a teacher's classroom to the time they leave. That measured growth, so it's argued, can be used to quantify and determine a teacher's purported effect on student growth in achievement over time.

Five sticky issues

In practice, however, whether these models are working as intended is under debate. Five issues are at the core of the disputes surrounding VAMs:

- *Reliability.* Teachers classified as effective one year might be classified as ineffective the next, or vice versa, and often to the extremes. Given the stability of teachers' levels of effectiveness otherwise, these swings should not occur.
- *Validity.* There's no evidence that indicators of teacher value-added are adequately correlated with at least one other concurrent measure of teacher effectiveness, such as supervisors' observational assessment of teachers or students' survey-based assessment.
- *Bias.* Evidence suggests that teacher value-added estimates systematically differ, given the varying demographic characteristics of students nonrandomly assigned to their classrooms. This occurs despite the statistical controls put in place to block bias.
- *Transparency.* Teachers and administrators don't seem to understand the models being used to evaluate them, which simultaneously thwarts the extent to which they might use their value-added estimates to improve instruction or initiate reforms.
- *Fairness.* Current research suggests that predominantly math teachers and teachers of reading and language arts are being held accountable using these systems, leaving about 70% of all public school teachers value-added ineligible. The ineligible teachers typically teach children in early childhood and high school grades and in noncore subject areas such as social studies, science, art, music, and physical education.

VAMS and high-stakes decision making

Consequential use of VAMs to make high-stakes decisions — such as promotions, tenure, merit pay, or termination — is also a major area of concern. For example, although research suggests that, ideally, three years of teacher-level data are needed to make the most accurate value-added estimates, some states' tenure and due process laws have provisions that allow districts to terminate or untenure teachers using only one or two consecutively unsatisfactory value-added scores (for example, Delaware, Florida, Indiana, and Pennsylvania). Elsewhere, in the Houston Independent School District, 221 teachers were terminated because they demonstrated "insufficient student academic growth reflected by [their] value-added scores" (Amrein-Beardsley et al., 2016). In New York, Sheri Lederman, an 18-year veteran teacher who, by all other accounts, is an excellent teacher, received an "ineffective" rating, or, more specifically, a score of 1 out of 20 one year after receiving a 14 out of 20 (Harris, 2016). Lederman successfully sued the state, after which the state retracted her value-added score and the erroneous rating. An additional 14 VAM-related lawsuits are ongoing across the United States (Education Week, 2015).

The unintended consequences

There are also unintended consequences that have gone unrecognized. These may include increased competition, increased isolation, heightened frustration, decreased morale, and diminished trust among educators. Research here is scant and often dismissed as anecdotal; a greater focus on research in this area would help capture both the intended and unintended consequences of implementing VAM.

Reaching out to the experts

It's crucial to clarify what those with expert knowledge of value-added models have to say about those models and their use. Accordingly, researchers collected the opinions of a subset of scholars who have researched and written about VAMs in some of the most prestigious research journals the field of education has to offer. Researchers surveyed 67 authors of 28 articles published before 2015 on the topic of value-added models. (A list of authors, institutional affiliations, and name of journal is available in an expanded version of this article at http://bit.ly/2bm3Pzc). We received responses from 33 of the authors we contacted. Here's what they reported.

Defining the value in value-added

Researchers first asked expert authors what "value" they thought VAMs add or could add in education. Respondents most often answered with a series of standard definitions that capture the theories behind VAMs. Respondents wrote, for example, that VAMs offer "one standardized indicator of teachers' contribution to student growth in math or [reading/language arts]"; that VAMs "simply [offer] a way to quantify how much students have learned"; and that "a properly specified VAM [provides] the closest thing we have to [a] causal estimate of the impact of a school or teacher on student achievement as measured by state tests."

Some also positioned VAMs against what they deemed a set of subpar alternatives. These respondents said that VAMs are better than alternatives, such as achievement or "snapshot" indicators or "subjective" observational systems, because VAMs "more objectively" provide "information that is external from the [school] system about student performance" over time from a growth or "longitudinal perspective." Such a perspective "is moving in the right direction" as "no other type of performance measure — classroom observations, student surveys, or subjective evaluations [of teachers] by principals — has been shown to be an unbiased predictor of contributions to achievement growth."

Other respondents, however, were most critical of these theories and assumptions, commenting that VAMs have serious limitations. They wrote that VAMs "are, if nothing else, one imperfect measure (among many) of the work that teachers do" and "whether VAMs can be used to evaluate teachers is doubtful, knowing what we already know from the research" about them.

Second, respondents most commonly identified VAMs as being a beneficial diagnostic or formative tool, noting that the models could "add value" to education, given their perceived abilities to identify low- and high-performing teachers. VAMs also could serve as a formative indicator for teachers. For example, one respondent said "VAMs can be a useful metric, especially in the case where teachers are doing consistently poorly and year after year, their students are not making progress." Respondents indicated that VAMs should identify outlier teachers, schools, and classrooms, which educators could study as positive (or negative) examples.

Third, respondents supported the use of VAMs as a general, large-scale research and evaluation tool,

Join the conversation

facebook.com/pdkintl
@pdkintl

The models could "add value" to education, given their perceived abilities to identify low- and high-performing teachers.

indicating that VAMs can help evaluate "programs, policies, and practices" and "can be a useful analytical tool in identifying important impacts in education [as per] interventions and/or practices."

Last, respondents described how VAMs might add value if used as part of evaluation systems based on multiple measures. These respondents specified that if VAMs were to be used as a high-stakes accountability indicator, VAM output would have to be viewed in conjunction with other correlated, evaluative indicators "at minimum." In other words, VAMs do "provide a method of measuring teacher effectiveness, which, combined with other methods, could improve the evaluation of teachers." However, all methods used to evaluate teachers should be "considered and understood by all involved," allowing "the discussion of VAMs [to lead] to a more rigorous assessment of teachers through [and while not dismissing] other methods."

Concerns about reliability and validity

The broadest concern respondents raised had to do with the classic methodological issues with VAMs and, more specifically, with VAMs' levels of reliability and evidences of validity. In terms of reliability, respondents mentioned the lack of consistency in teacher-level VAM estimates over time, which manifests itself in the "estimates [being] highly inconsistent" and "the error rate [being] too high [when] determining good and poor performers." VAM-based estimates are "noisy in that they fluctuate from year to year."

In terms of validity, respondents doubted that VAM estimates actually can yield the "accurate indicators of educational quality" for which they are meant, especially at the inferential level. Concerns included how VAM estimates don't seem to line up well (i.e., correlate) with other similar indicators of teacher effectiveness. Respondents also pointed to how VAMs are open to distortions themselves — such as student sorting, teaching to the test, cheating, and artificial score inflation — just like other methods of evaluating teacher effectiveness that are often deemed less "objective." This "has to do with Campbell's law . . . in which substantial efforts are made to game tests and [VAM] systems as they become increasingly important [and consequential] components of evaluation systems."

Concerns about bias

Bias caused by the nonrandom placement of students into teachers' classes, or by teachers being consistently assigned more "extreme" sets of students with "extreme" sets of group demographics, also raised respondents' concerns. Respondents wondered about "the ways in which the teaching of

student subpopulations gets to be analyzed," especially when "so many researchers and VAM-oriented minds put [diverse] learners in very homogeneous categories and boxes for [VAM] analyses." In other words, of great concern is whether the types of students teachers teach (e.g., placed into their classrooms via multiple, nonrandom class assignment methods) can positively or negatively influence teachers' value-added estimates. Although in an ideal world unbiased estimates can be obtained through experimental variation in students assigned to teachers, in reality "randomizing students into classrooms would be impossible" (see also Paufler & Amrein-Beardsley, 2014). Likewise, failing "to fully [e.g., statistically] account for the external factors (e.g., student levels of poverty, race and ethnicity, attendance) in such models could systematically bias teachers and schools that [primarily] educate underserved student populations." This, too, has major implications for education policy and practice.

Respondents also were concerned with how VAM estimates are sometimes used, misused, and abused in high-stakes decision making. VAMs "seem to be used in very different ways by different groups." For example, some use VAM estimates for informational purposes (e.g., to analyze individual or general trends over time), yet others use VAM estimates for more extreme, consequential purposes (e.g., the red flagging of teachers' professional files when teachers are deemed to have relatively lower value or "value-added"). These ways have "important implications," as this "can lead to capricious systems where teachers are evaluated based on flawed ratings." Moreover, respondents were concerned with "school and district administrators leaning too heavily on VAMs to make decisions regarding termination, raises, etc., or school funding decisions being tied too closely to these measures." "Even if the parameter estimates are reasonable," noted one respondent, "it is generally not appropriate to use them for unwarranted decisions."

Respondents cautioned that "the teacher represents only a fraction of the variability in student test scores (about 10%)"; hence, "assigning causality is problematic based on VAMs" in that "it's hard to achieve causality for many reasons, some of which can be mitigated through statistics and some of which cannot." The focus on the causal interpretation of VAMs also has "frustrated [public school teachers] since [the causal link to be made] suggests that all of the problems with education quality stem from teachers."

Concerns about transparency and fairness

Next on the list were transparency and fairness, with respondents noting that "school administrators and teachers are most affected by VAMs, especially in terms of high-stakes consequences." Yet administrators and teachers "do not [apparently] understand VAMs," given that the VAMs used to measure and evaluate them are "nontransparent," when VAM-based estimates, their purposes, implications, and mechanisms behind them are not easily understood by the average users (e.g., teachers or administrators). Defining transparency as the extent to which something is easily seen and readily capable of being understood, VAM-based estimates must be made transparent in order to be understood so they can ultimately be used to "inform" change, growth, and hopefully future progress in "formative" ways. Also, "the public has not been adequately educated on how to interpret [VAM] . . . scores."

Fairness is an issue because VAM scores "are not available for many teachers" and the "percentage of educators for whom VAMs currently cannot be applied is rather high." This is because the "student outcome measure . . . cannot be used for the majority of teachers ([in] untested grades and subjects)." Hence, few teachers are eligible to be held accountable using VAMs. Respondents collectively asserted that "teachers also contribute to other measures (i.e., noncognitive outcomes)" but that "most VAMs use basic skill-type tests . . . based only on [reading/language arts] and math achievement." These tests "are limited in many ways, so they do not give us a full picture of how well teachers are improving achievement" because they "may not [appropriately] reflect appropriate student learning."

It's also worth noting that the respondents' collective opinions reported here also align with recent position statements on value-added models released by the American Statistical Association (2014) and the American Educational Research Association (2015).

What about high-stakes consequences?

Most respondents said high-stakes consequences might be attached to VAM-based estimates but only under certain conditions, given certain caveats, and "depend[ing] on the decisions" to be made. However, these respondents didn't see the need for "throwing out VAMs" nor did they "see a problem with bringing all the evidence to bear on the situation," believing that if VAMs were used along with other indicators of teacher quality — and, especially, if VAMs and other indicators didn't contradict themselves — consequential decisions attached to multiple indicators might be warranted: "If teachers are consistently (for multiple years) receiving low value-added estimates, then that information should be considered in personnel decisions, along with other measures of teaching quality

Teachers classified as effective one year might be classified as ineffective the next, or vice versa, and often to the extremes.

(observational measures, portfolio measures, etc.)" that "more heavily weight factors in direct control of the teacher (like observation scores)." But there's a caveat: "Each of the [additional] components [would need to] meet criteria of suitable reliability (precision), validity (accuracy), and financial plausibility." Hence, VAMs "might be used to build a hypothesis, but then other information should be used to test the hypothesis" in the most "scientific" and "judicious" way possible, keeping in mind that "complex, real-life situations demand a much more complex analysis . . . which needs to always include a degree of human analysis" and judgment.

Only two respondents said attaching consequences to VAM estimates was currently possible; all of the other respondents said the opposite, regardless of the caveats. Although respondents agreed that decisions about teacher effectiveness shouldn't be "devoid of data," they said "teachers and administrators at local sites should be the ones who decide what [VAMs] are worth . . . provided they are aware of and fully understand [VAM] estimates' limitations so that they can critically and wisely use them . . . or not."

Getting smarter about value-added

Value-added models are hard to oppose: "It's very difficult to counter a measure that supposedly captures a teacher's contribution to student academic growth and that uses 'objective' data to do so." Likewise, VAMs "are little understood by policy makers, educators, education leaders, and the general public," creating an ongoing political struggle difficult to win. This makes it all the more important for education policy makers and leaders at the federal, state, and local levels to listen to the experts as a whole not just to those representing the minority view. It may even be true, as noted one expert respondent (Darling-Hammond, 2015; Darling-Hammond, Wilhoit, & Pittenger, 2014) in this study, that "[VAMs] are not a solution at all." We may well need a different expert-based solution.

VAMs have become an important part of the education research and policy landscape, and researchers continue to draw different conclusions about them. It's more important than ever to build spaces in which we can construct and share the collective knowledge, intelligence, wisdom, and expertise of those (with the possible exception of teachers, perhaps) who understand VAMs the best. We hope readers will walk away from this study with a good bit of this shared knowledge, intelligence, wisdom, and expertise. K

References

American Educational Research Association (AERA). (2015). *AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs.* http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html

American Statistical Association (ASA). (2014). *ASA statement on using value-added models for educational assessment.* www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A., Collins, C., Holloway-Libell, J., & Paufler, N.A. (2016, January 5). Everything is bigger (and badder) in Texas: Houston's teacher value-added system. *Teachers College Record.* www.tcrecord.org/Content.asp?ContentId=18983

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? *Educational Researcher, 44* (2), 132-137.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives, 22* (1), 1-86. http://epaa.asu.edu/ojs/article/view/1724

Education Week. (2015, October 6). Teacher evaluation heads to the courts. *Education Week.* www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Harris, E.A. (2016, May 10). Court vacates Long Island teacher's evaluation tied to test scores. *New York Times.* www.nytimes.com/2016/05/11/nyregion/court-vacates-long-island-teachers-evaluation-tied-to-student-test-scores.html?_r=0

Paufler, N.A. & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal, 51* (2), 328-362.

Author	Institution	Journal	Publication Year	Citations
Algina	University of Florida	JEBS	2004	150
Amrein-Beardsley	Arizona State University	ER, AERJ	2008, 2014	140, 2
Ariet	University of Florida	JEBS	2004	150
Ballou	Vanderbilt University	JEBS	2004	397
Briggs	University of Colorado, Boulder	JEBS, JEBS	2011, 2013	11, 1
Broatch	Arizona State University	JEBS	2012	6
Brown	Stanford University	ER	2014	0
Buzick	ETS	ER	2013	11
Carter	University of Buffalo	JEBS	2004	150
Castellano	University of California, Berkeley	JEBS	2014	0
Chiang	Mathematica	JEBS	2013	6
Cohen	University of Virginia	ER	2014	0
Cowen	University of Kentucky	ER	2013	12
Domingue	University of Colorado, Boulder	JEBS	2013	1
Fisher	Fisher Education Consulting	JEBS	2004	150
Fox	Stanford University	EEPA	2014	3
Goldhaber	University of Washington	EEPA	2013	10
Goldschmidt	CalState Northridge	EEPA	2013	10
Grossman	Stanford University	ER	2014	0
Hamilton	RAND	JEBS	2004	393
Harris	Tulane University	AERJ	2014	4
Hill	Harvard University	AERJ	2011	87
Ingle	University of Louisville	AERJ	2014	4
Jacob	University of Michigan	EEPA	2011	17
Jones	ETS	ER	2013	11
Kapitula	Calvin College	AERJ	2011	87
Karl	Adsurgo LLC	JEBS	2013	2
Kelly	University of Notre Dame	ER	2007	29
Koretz	Harvard University	JEBS	2004	393
Kupermintz	University of Haifa, Israel	EEPA	2003	184
Lefgren	Brigham Young University	EEPA	2012	10
Lockwood	RAND	JEBS	2004	393
Loeb	Stanford University	EEPA	2014	3
Lohr	Arizona State University, Westat	JEBS, JEBS	2012, 2013	6, 2
Louis	Johns Hopkins University	JEBS	2004	393
Lucas	Alachua County School Board	JEBS	2004	150
Ma	University of Florida	JEBS	2004	150
Martineau	Michigan Department of Education	JEBS	2006	106
McCaffrey	RAND	JEBS	2004	393
Monczunski	Purdue University	ER	2007	29
Papay	Harvard University	AERJ	2010	98
Paufler	Arizona State University	AERJ	2014	2
Polikoff	University of Southern California	EEPA	2014	1
Porter	University of Pennsylvania	EEPA	2014	1
Rabe-Hesketh	University of California, Berkeley	JEBS	2014	0

Author	Institution	Journal	Publication Year	Citations
Raudenbush	University of Michigan	JEBS	2004	197
Reckase	Michigan State University	JEBS	2004	39
Resnick	University Florida	JEBS	2004	150
Ronfeldt	University of Michigan	ER	2014	0
Roth	University of Florida	JEBS	2004	150
Rubin	Harvard University	JEBS	2004	209
Rutledge	Florida State University	AERJ	2014	4
Sanders	SAS Institute	JEBS	2005	397
Schochet	Mathematica	JEBS	2013	6
Sims	Brigham Young University	EEPA	2012	10
Skrondal	Norwegian Institute of Public Health	JEBS	2014	0
Soland	Stanford University	EEPA	2014	3
Stuart	Harvard University	JEBS	2004	209
Tekwe	Johns Hopkins University	JEBS	2004	150
Tseng	Berkeley Policy Associates	EEPA	2013	10
Turkan	ETS	ER	2013	11
Umland	University of New Mexico	AERJ	2011	87
Weeks	University of Colorado, Boulder	JEBS	2011	11
Winters	University of Colorado, Colorado Springs	ER	2013	12
Wright	SAS Institute	JEBS	2005	397
Yang	Arizona State University	JEBS	2013	2
Zanutto	University of Pennsylvania	JEBS	2004	209