Understanding the Basic Problems with the Use of Value Added in Ohio

Randy L. Hoover, PhD

(2014)

Value added is unquestionably the most complex element of Ohio's accountability metrics machine. Arguably, it is the most controversial and often contentious element in any of the state accountability systems using it. As a key requirement of Obama's Race to the Top (RttT), value-added models (VAM) have become the primary high-stakes metrics by which educators are being judged for their effectiveness. Having studied and researched VAM for a number of years, I will say without hesitation that value added as it is used in high-stakes education accountability systems represents a truly remarkable example of politicians and policy makers ignoring a vast amount of good research in favor of ideological and special-interest group motivations. In spite of an overwhelming amount of quality research showing value added to be unreliable and inappropriate for use in determining teacher effectiveness, it stands as the lynchpin of Ohio's teacher evaluation system.

Understanding the critical issues and research arguments against the use of value added is an extremely important career necessity given that value added has such extremely high-stakes consequences for public school teachers and administrators. As is the case with any test or metric used to evaluate students, teachers, or schools, critical examination of the test or metric itself is necessary as well as is *critical examination of the context within which the test or metric it is used.*

It is not the intent of this summary paper to be a literature review of research on value added. Nor is it intended to probe the nature of value-added calculus in and of itself. *The intent herein is primarily to give a big-picture view of the real-world context of the use of value added as the high-stakes threat to the professional integrity of honest, hardworking, and intelligent classroom teachers.* However, I do strongly recommend examining some of the works listed on the website[1] of Audrey Amrein-Beardsley, PhD at Arizona State University. Beardsley is a highly recognized and respected expert in the area of value-added metrics and their use in educational accountability systems. Similarly, for those looking for a succinct expose' of VAM that is very readable and well researched, I recommend Amrein-Beardsley's book "Rethinking Value-Added Models in Education" published by Routledge in spring of 2014.

To address the contextual issues of educational accountability systems in terms of how the tests and procedures are used in those systems, the *Standards for Educational Accountability Systems* (2002) were produced by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) by a consortium comprised of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

The CRESST standards consist of 22 criteria that represent rigorous guidelines for all educational accountability systems. The test standards are widely recognized as the most authoritative statement of professional consensus regarding expectations on matters of validity, fairness, and other technical characteristics of tests. Adherence to all of the 22 standards is a requirement for any given accountability system to be fully warranted and therefore credible.

The first CRESST standard violated by the state deals with value added and the issue of test validity. Although the state attempts to document that the standardized achievement tests are valid, the documentation only addresses validity for the use of the tests for assessing student achievement. Because the psychometric definition of test validity is tightly restricted to the one specific use of the test, any other use of the test must be precisely documented separate from the first-use documentation. *The documented claim for achievement test validity is restricted to the use of the tests as measures of student achievement and no other use.* But the tests are also used in Ohio for measuring teacher effectiveness

[1] See: http://vamboozled.com/recommended-reading/value-added-models/

through value added. Thus, this is a second specified use for the achievement tests for which there is no clear validity documentation. CRESST Standard 9 addresses validity for more than one test purpose:

> 9. The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.
>
> Comment: Validity is dependent on the specific uses and interpretations of test scores. It is inappropriate to assume that a test that is valid when used for one purpose will also be valid for other uses or interpretations. *Hence, validity needs to be specifically evaluated and documented for each purpose*. (p.9) (Italics added.)

The validity of value-added metrics is documented by Education Value Added Assessment Systems (EVAAS), the corporate vendor thoroughly, and at great length. However, the documentation of value added by SAS addresses the metrics in and of themselves. And as said before, value added in and of itself is neither good nor bad. Indeed, it has a seemingly distinguished history of use in agriculture and industry. What is not done in the documentation is to show the achievement tests to be valid for use as data sources for determining teacher effectiveness.

Of course, the best argument against value added use comes from the fact that it is based entirely on invalid standardized test scores *in the first place*. I once conversed with two high-level EVAAS experts in psychometrics about the issue of Ohio's standardized test validity. They both candidly admitted that the value added model used in Ohio operates *on the assumption* that the tests are fully valid. I had a similar conversation with a high-level employee of *Battelle for Kids,* who answered similarly and specifically told me with great concern that the way Ohio is using value added is inappropriate to the basic methodological requirements.

Subsequent to these conversations, in the late winter of 2014, Jamie Meade, Battelle Managing Director of Strategic Measures at Battelle for Kids, published a blog statement entitled "Value Added Data Quality: Too Often it is in the Eyes of the Beholder[2]" that is a clear attempt to distance Battelle for Kids from what is being done with the way Ohio is using value added. Given the politics and money involved across the tripartite relationships among Battelle for Kids, EVAAS, and ODE, it is not surprising that the blog entry is very delicately written. I think it is important to note that the Battelle Memorial Institute, founded in 1929, has a very distinguished history of doing research and development for science, business, industry, and government. Battelle for Kids, however, is a non-profit group formed by the Ohio Business Roundtable with a grant from Battelle Memorial Institute. It is not a subsidiary of Battelle Memorial Institute. The central and dominant hand in Battelle for Kids being a ruse for selling value added in Ohio is the Ohio Business Roundtable. At the national level, The Business Roundtable is a right-wing special-interest group known for being the primary special-interest group in support of anti-public school legislation including NCLB, vouchers, charter schools, and pseudo accountability models.

Returning to the issue of Ohio's blatant violations of key standards for holding educators accountable for student performance, we turn to Standard 15:

> 15. *Stakes for accountability systems should apply to adults and students and should be coordinated to support system goals.*
>
> Comment: Asymmetry in stakes may have undesirable consequences, both perceived and real. For example, *if teachers and administrators are held accountable for student achievement, but students are not, then there are likely to be concerns about the degree to which students put forth their best effort in taking the tests.* Conversely, it may be unfair to hold students accountable for performance on a test without having some assurance that teachers and other adults are being held accountable for providing students with adequate opportunity to learn the material that is

[2] See the blog entry at http://blog.battelleforkids.org/blog/2014/4/30/value-added-data-quality-too-often-its-in-the-eyes-of-the-be.html

> tested. Incentives and sanctions that push in opposite directions for adults and for students can be counterproductive. They need to be consistent with each other and with the goals of the system. (p. 4) (Italics added.)

Because value added has extremely high-stakes ramifications, the effects of asymmetry can be disastrous for teachers. The problem of asymmetry is a context problem not inherent in value added metrics themselves. The problem, irrespective of value added metrics, is in the way Ohio chooses to conduct value added data collection via the standardized tests. Ohio makes the assumption that all students have identical, positive attitudes toward school and their teachers such that all students whose tests are used as the data for value added will try as hard as they can to do as well as they can on standardized test destined to factor into the teachers value added score. Any students in a given class who have a capricious attitude toward the class, the teacher, the school, or the test, significantly affect the particular teacher's value added score. Because of this, the test scores are not worthy of inclusion for the teacher's evaluation rating.

To further exemplify Ohio's capricious attitude toward its teachers, Standard 11 of CRESST requires careful reporting of any and all possible sources of threats to accuracy in classification for high stakes rewards or sanctions. In the case above, not only does Ohio ignore the significant effects of asymmetry, Ohio also ignores reporting that it ignores Standard 14 by also ignoring Standard 11:

> 11. If test data are used as a basis for rewards or sanctions, evidence of the technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published.
>
> Comment: Because tests are fallible measures, classification errors are inevitable when tests are used to classify students or institutions into categories associated with rewards or sanctions. In order to judge whether the risk of errors is acceptably low, *it is essential that information be provided about the probability of misclassifications of various kinds.* (p. 3) (Italics added.)

With value added, it is at the root level of the standardized tests where the misclassifications begin and are subsequently compounded through the assumptions and metrics of value added itself. Most simply, the problem exists in terms of the test score confidence intervals. The confidence interval for a test score is *the calculated range of the score, not a precise single value as is suggested by the accountability system.* Whether the tests are no-stakes, low-stakes, or high-stakes, ignoring confidence intervals (range of the score) makes a tremendous difference. The higher the stakes, the more significant misclassifications become. In the case of Ohio, where value added is a major element of teacher evaluation, the stakes are about as high as they can be. Typically, test scores like those in the example *are* the primary sources of data for high-stakes categorization of teachers, individual schools, and districts.[3]

When student test scores are used as data for overall accountability systems, the likelihood of precise meaning is impossible mathematically. Value-added models are the epitome of fuzziness when it comes to precision. Mathew Di Carlo, senior research fellow at the Albert Shanker Institute, writing about the work of Sean Corcoran (2010) of the Annenberg Institute for School Reform, notes how using value-added models (VAM) greatly *amplifies* the role of measurement error:

> Interpreting a teacher's VAM score without examining the error margin is, in many respects, meaningless. For instance, a recent analysis of VAM scores in New York City shows that the *average* error margin is plus or minus 30 percentile points. That puts the "true score" (which we can't know) of a 50th percentile teacher at somewhere between the 20th and 80th percentile—an incredible 60-point spread (though, to be fair, the "true score" is much

[3] In my 2000 study, when I controlled for the lived experience factor (socioeconomic status), I found many apparently high-performing districts to be performing well below their state ratings and many low-performing districts actually performing far above.

> more likely to be 50th percentile than 20th or 80th, and many individual teacher's error margins are less wide than the average). If evaluation systems don't pay any attention to the margin of error, the estimate is little more than a good guess (and often not a very good one at that). Now, here's the problem: Many, if not most, teacher evaluation systems that include VAM—current, enacted, or under consideration—*completely ignore this*. (Di Carlo, 2010)

Even if there were only a relatively small spread, the extreme high-stakes nature of the Ohio Teacher Evaluation System demands that the chances for misclassification be made explicit and public. Ignoring Standard 11 and also Standard 16 (below) is enormously problematic for Ohio teacher in terms of the value-added side of the evaluation system. The teacher is at the mercy of value added amplifying measurement error and has no way of appealing the victimization it can cause.

Likewise and as discussed in the OTES paper on this site, given that context problems clearly exist in the data collection for value added as identified here, the fundamental principles of equity in our democratic society demand there be some form of appeals process for adjudicating value-added misclassification. CRESST Standard 16:

> 16. Appeal procedures should be available to contest rewards and sanctions.
>
> Comment: Extenuating circumstances may call the validity of results into question. For example, a disturbance during test administration may invalidate the test results. *Also, individuals may have information that leads to conflicting conclusions about performance. Appeal procedures allow for such additional information to be brought to bear on a decision and thereby enhance its validity.* (p. 4) (Italics added.)

One of the most essential principles of American justice is known as *due process*. This legal principle flows from the 14th amendment to the U.S. Constitution as insurance against arbitrarily placing a person in harms way or depriving a person of a property right in an unfair manner. Few people today realize that the reason for granting tenure to teachers was instituted to guarantee that once a teacher demonstrated their full professionalism, they then were entitled to due process before they could be fired. *Tenure did not and never did guarantee teachers a job. Tenure only guaranteed teachers the Constitutional right to due process before being removed from their job.* Though the high-stakes rating resulting from value added scores may not always mean being dismissed, the value of affording due process to the teacher is to insure there is integrity in the accountability system. In large part, this is what motivates the CRESST Standard 16.

Another very critical flaw in value added is the demonstrably false assumption that there is a one-to-one correspondence between the teacher and student achievement. As similarly noted in the OTES paper on this site, value-added metrics assume that teachers are the direct cause of student scores—that a specific teacher's effects are revealed directly and fully by standardized test scores. Value added *assumes* that teacher effects are significant, specific, and precisely quantifiable when they are not. Richard Rothstein (2010), writing for the Economic Policy Institute (EPI), speaks directly to this entirely false value-added assumption:

> It has become conventional in educational policy discussion to assert that "research shows" that "teachers are the most important influence on student achievement," There is, in fact, no serious research that shows any such thing. The assertion results from a careless glide from "teachers being the most important in-school influence" to teachers being the most important influence overall. But because school effects on average levels of achievement are smaller than the effects of families and communities, even if teachers were the largest school effect, they would not be a very big portion of the overall effect. A child with an average teacher who comes from a literate, economically secure, and stable family environment will, on average, have better achievement than a child with a superior teacher but with none of these contextual advantages.

Addressing the same problem, Scott McLeod (2013) writes:

> Another issue worth noting is that even if teacher effects could be teased out, decades of peer-reviewed research show that teachers only account for about 10% of overall student achievement (give or take a few percentage points). Another 10% or so is attributable to other school factors such as leadership, resources, and peer influences. The remaining 80% of overall student achievement is attributable to non-school factors such as individual, family, and neighborhood characteristics. A few exceptional 'beating the odds' schools aside, these ratios have remained fairly stable (i.e., within a few percentage points) since they were first noted by the famous Coleman Report of the 1960s. Given the overwhelming percentage of student learning outcomes that is attributable to non-teacher factors, it is neither ethical nor legally defensible to base teacher evaluations on factors outside of their control. (McLeod, 2013)

Both Rothstein and McCloud not only support the findings of my own research findings on Ohio district performance (Hoover, 2000, 2008) as to the limited extent to which teachers can have academic impact on student achievement as measured by the tests, but McLeod nearly echoes word for word the essence of my definition of pseudo accountability in his last sentence.

To fully grasp the ramifications of value added, we need to realize that the value-added metrics machine is designed to further lock public school teachers into the world of pseudo accountability, a world where they are held responsible for phenomena completely beyond their control . It does so with the highest stakes for teachers of any element of the accountability model. *To be dismissed from a teaching job because of value added scores is to be dismissed for reasons entirely beyond the professional control of that teacher irrespective of the actual quality or effectiveness of the teacher*.

To this point, it is primarily context issues, issues about how value added is used, that form arguments against its use. These issues are so serious that value added merits the strongest opposition possible from any and all groups in Ohio that are seeking authentic accountability as a model for teacher evaluation. We must also remember that Ohio's system for evaluating principals is equally flawed because the value added effects that stem from the entire range of context and metrics issues. Likewise, if we are to have credible school performance reports for the general public, these major failures in how value added is used must be rectified.

There are also elements within the metrics of value added that also serve as evidence for the argument that it should not be used as a measure of effectiveness. For example, the number of years included in calculating teacher value-added scores is extremely important mathematically. We know that it is indefensible mathematically to use fewer than three years of performance-score data with many critics arguing that five years is likely the minimum for any accuracy at all. It is important to know that this debate ignores all other arguments against value added use.

The last problem is that the complete value-added algorithm is unavailable for public examination because it is proprietary knowledge held by EVAAS, the corporation responsible for administering and providing value-added reports for a number of states and home to William Sanders, the creator of the value-added application to educational evaluation. For educators to be held accountable through a mechanism that they are not permitted to fully scrutinize is a democratic outrage that speaks volumes to the hidden agendas of so-called education reform.

There is no question whatsoever that teacher evaluations based in value added place our classroom teachers into job-dependent situations over which they hold no control. We simply cannot allow Ohio's classroom teachers to be victimized by a demonstrably indefensible system. That our major professional associations have stood by idly, failing to inform and educate their members to the threat, is indeed a sad commentary on teacher advocacy and the well being of Ohio's public school teachers.

References

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability.* New York, NY: Routledge.

Center for Research on Evaluation, Standards, and Student Testing (2002). *Standards for educational accountability systems*. Los Angeles: UCLA.

Corcoran, S. (2010). Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice. Retrieved from http://www.scribd.com/doc/37648467/The-Use-of-Value-Added-Measures-of-Teacher-Effectiveness-in-Policy-and-Practice#download

Di Carlo, M. (2010). The war on error. Retrieved from http://shankerblog.org/?p=1383

Hoover, R. (2000). *Forces and factors affecting Ohio proficiency test performance: A study of 593 Ohio school districts*. Retrieved from http://people.ysu.edu/~rlhoover/OAT-OGT/index.html

Hoover, R. (2004). *(Re)understanding educator accountability: Pseudo vs. authentic accountability.* Retrieved from http://people.ysu.edu/~rlhoover/OAT-OGT/index.html

Hoover, R. (2008). *A Re-examination of forces and factors affecting Ohio school district OAT and OGT performance.* Retrieved from http://people.ysu.edu/~rlhoover/OAT-OGT/index.html

McLeod, S. (2013). *Value added measures: VAM*. Retrieved from http://dangerouslyirrelevant.org/resources/value-added-measures.

Meade, J. (2014). *Value Added Data Quality: Too Often it is in the Eyes of the Beholder.* Retrieved from http://blog.battelleforkids.org/blog/2014/4/30/value-added-data-quality-too-often-its-in-the-eyes-of-the-be.html

Rothstein, R. (2010). *How to fix our schools: It's more complicated, and more work, than the Klein-Rhee 'manifesto' wants you to believe*. Economic Policy Institute, Issue Brief #286. Washington, DC: EPI