Deconstructing Louisiana's Legislative Report on Value-Added Modeling

Wade Smith & Adam C. Elder
Southeastern Louisiana University
500 W University Ave, Hammond, LA 70402, United States

Abstract

As of 2019, thirty-four states were requiring an objective measure of student growth as part of their teacher evaluation efforts (National Council on Teacher Quality, 2019). Louisiana is one of those states and state law requires a report to be given to the legislature detailing the progress of Value-Added Modeling (VAM) for teacher evaluations. This paper focuses on the information that can and cannot be found in the legislative report. Additionally, the research reports on the conclusions derived from the information provided and their limitations. The full paper documents several shortcomings considered essential to have a complete understanding of the actual state of affairs for VAM in Louisiana. The shortcomings include the use of methodologies that are known to introduce noise into the results, the lack of critical information necessary to evaluate the reliability and stability of the VAM model, and the amount of variance explained by the model. The findings discussed in this paper have important implications for VAM policy and research in Louisiana specifically as well as in states that use VAM generally.

Keywords: Value added models, policy analysis, teacher evaluation

The rise of sophisticated statistical models and the push for strong and reliable teacher evaluations by policy makers paved the way for Value-Added Modeling (VAM; Haertel, 2013).The move towards a statistical model to identify and isolate teacher effects has been underway now for nearly 25 years, beginning with Sanders and Rivers' (1996) initial attempt towards this end. The prospect of developing a statistical model that can isolate for teacher effects was accelerated during the Obama administration's Race to the Top (RTTP) program. States with successful RTTP applications committed to participating in the process of using systems of educator evaluation that linked to student achievement growth measures (U.S. Department of Education, 2009). Forty-six states and the District of Columbia applied for RTTP funds and 19 states received over four billion dollars in federal grants. Altogether, 34 states revised their laws to align them with RTTP expectations (Obama White House Archives, n.d.).

Louisiana is one of the states that legislated VAM in 2010 and began implementation in 2012. Originally, the results of VAM were prescribed by law to account for 50% of a teacher's evaluation. In 2017, the sponsor of the legislation to incorporate VAM into teacher evaluations changed course, filing a bill to do away with VAM inclusion in teacher evaluations because of his doubts regarding the reliability and validity of the results (Sentell, 2017). The bill was killed in committee (Arsement, 2018). In Louisiana, fourth through eighth grade teachers receive value-added scores for English Language Arts, Math, Science, and Social Studies. Instructors for Algebra I and Geometry are also part of Louisiana's VAM (Louisiana Department of Education [LDOE], n.d.). Teachers participating in VAM have 35% of their evaluation based upon their value-added scores, 15% upon students reaching their learning targets, and 50% upon professional practices. Since 50% of the evaluation involves some iteration of local or state derived growth targets, it is clear that focus upon student learning is a priority.

Certainly, there is nothing controversial about teachers strongly emphasizing student learning. Although the concept is clear, the means of measuring teacher effects on student learning are not without problems. "Specifically, teacher effects based on statistical estimates may actually represent the combined contributions of many factors in addition to the real teacher contribution we are after" (Braun, 2005, p. 8). The difficulty in separating combined contributions of many factors from the real teacher contribution is germane to any evaluation that relies heavily upon VAM and is one of the factors contributing to an expanding conversation about VAM's validity.

VAM Concerns

A critical assumption in VAM is that any effect not captured by the VAM equation will be captured by the student fixed-effect components (Braun et al., 2010). Braun and his colleagues, in that same paper, noted that non-random processes such as scheduling of students to a particular classroom are not controlled by student fixed effects. This finding brings into question underlying assumptions for VAM. In 2005, Braun made the case:

According to statistical theory, the ideal setting for obtaining proper estimates of teacher effectiveness ... is a school system in which, for each grade, students are randomly grouped into classes, and teachers in that grade are randomly allocated to those classes. Roughly speaking, randomization levels the playing field for all teachers in that each teacher has an equal chance of being assigned to any class.

The data generated in such a setting would allow us to obtain a reasonable estimate of each teacher's effectiveness, as well as a measure of the precision to be attached to the estimate. A finding that the average student growth associated with a particular teacher is significantly greater than the district average would be credible evidence for that teacher's relative effectiveness. (p. 8)

Anyone that has ever built a master schedule for a school is aware that the process is not random. If, as Braun asserts, randomization allows for a reasonable estimate of each teacher's effectiveness and the precision of that estimate, then there is reason to be concerned when randomized student selection is not realized. This sentiment was echoed by Paufler and Amrein-Beardsley (2014) who found that principals, teachers, and parents play a significant role in student placement decisions. Statistical efforts to control for randomization have been developed (Koedel& Betts, 2011), but these efforts are typically not applicable to large datasets such as those found in a state evaluation model.

Year-to-year variations in teacher scores are another threat to the validity of VAM results. In an effort to control for year-to-year variations in VAM results, Louisiana uses a three-year average of test scores. Although this practice is considered to mitigate for variation, there is evidence that suggests the opposite may be true. In North Carolina, for example, the number of teachers that shifted out of the bottom quintile to the top and vice versa actually increased over a three-year evaluation period (Goldhaber & Hansen, 2008).

Research on VAM has consistently found issues that should give pause to policymakers implementing VAMs for evaluative purposes. Studies have shown there are myriad concerns with the validity of VAM due to issues with model specification (e.g., Amrein-Beardsley, 2008; Goldhaber et al., 2013; Hill et al., 2011; Kersting et al., 2013; Schochet& Chiang, 2011) as well as the validity threats posed by other matters like test selection and timing (Papay, 2011). Bearing this body of research in mind, it is critical for state policies regarding VAM to be closely scrutinized.

VAM in Politics and Law

Nationally, there appears to be a flux in the use of, and sentiments towards, VAM. Alaska, Arkansas, Kansas, Kentucky, North Carolina, and Oklahoma have recently stopped requiring student-growth measures in formal evaluations. Connecticut, Nevada, and Utah have adopted a modified response, requiring some evidence of student learning while disallowing the use of standardized-test scores to create that evidence (Loewus, 2017). That said, many other states and Louisiana in particular, have invested considerable time and political capital into the realization of VAM as a substantial part of teacher evaluations. In the case of Louisiana, there has been reluctance to revisit VAM. Proponents claim the information provided by VAM is an important component of teacher evaluations and detractors mostly assert the move to VAM has taken place in a rushed fashion (Sentell, 2017).

By and large, the willingness to reconsider VAM has become a political issue, although the political arena is now being informed by legal actions as challenges to VAM reach the courts. A recent court ruling in *Houston Federation of Teachers et al. v. Houston Independent School District* (2017) determined that secret algorithms are incompatible with due process and denied the school system's request for summary judgement. Meanwhile, New Mexico is now wrestling with a similar proceeding (Amrein-Beardsley, 2018).Politics aside, the results of VAM efforts should speak for themselves. Are the consequences of VAM an objective way of looking at student success and a more accurate prediction of student scores as claimed by the LDOE as well as other states? Or is that expectation overly optimistic? Should VAM be included as a substantial factor with a fixed weight in teacher evaluation systems? Is the information generated by VAM sufficient to create confident evaluations?

To address these questions for Louisiana, a public records request was made to the LDOE for the following information: (a) de-identified datasets for VAM, (b) confidence intervals for the model, (c) variance explained by the model, (d) the regression equation used in the model, and (e) the stability of the model across years. In response to the records request, an emailed response from the LDOE indicated everything we were looking for would be found in the annual report entitled *Louisiana's Value-Added Assessment Model for Educator Evaluation and Support: A Report in Response to R.S. 17:3883(A)(8)*. The Department is required to provide this report to the House and Senate Committees on Education (LDOE, 2019). For brevity's sake, it will be referred to as the Report throughout the remainder of the paper. Given the import of the Report, it was appropriate to analyze the document to determine whether the information provided, in effect, told the complete story about VAM. A careful read of the Report left us with several questions that we were not able to answer and these questions were considered critical to having a proper perspective for the complete story.

Results of the Public Records Request

We received no written notice, per state law, that would exempt any of our requests from the data provided to us in our records request. Nevertheless, much of what we asked for was not to be found in the Report. Some of the requests are directly related to recommendations provided by the American Statistical Association's (ASA, 2014)

Expectations for the proper reporting of VAM results, including the use of confidence intervals or standard errors. Since we were not provided the requested information, and the Report was referenced by the LDOE as the source of all the information we requested, we used the Report as directed to finalize our analysis as described below.

Finding 1: Teacher ratings may not accurately reflect actual teacher effectiveness

The Report stated that teachers were assigned four levels of effectiveness based on concomitant percentiles predetermined by a committee prior to data analysis. The lowest 10% of the centralized database were labeled Ineffective, the next 39% were Effective: Emerging, the next 30% were Effective: Proficient, and the highest 20% were rated as Highly Effective, essentially creating a distribution that defines 50% of the teaching population as "below average" and 50% as "above average." We were unable to ascertain why these percentages and classifications were chosen. Why would a committee "know" that you have twice as many teachers that are highly effective compared to those who are ineffective before you have any data? It is not possible to know what "ineffective" looks like until the data are crunched. If that prediction could actually be made, then there would be no need for VAM—or any other form of evaluation for that matter. Letting the committee predict where every teacher would fall on the rating scale could provide the same results. At a minimum, the Report should provide a rationale for these predetermined categorizations.

Finding 2: There was considerable movement of teachers in the VAM ratings from year-to-year

Considerable movement of teachers' VAM ratings from year-to-year is a reality that has been previously documented in other studies (e.g., McCaffery et al., 2009) and Louisiana's data follows suit. Approximately one-fourth of the teachers rated in the bottom 10% maintained that status in the following school-year while another fourth were rated in the top half of all VAM-evaluated teachers. Nearly half of the "effective emerging" teachers moved classifications, with nearly 39% now being identified in the top half of teachers statewide. The majority of "effective proficient" teachers were reclassified the following school-year, with 39% newly identified in the bottom half of teachers statewide.

Many of the teachers impacted by VAM evaluations have seen their scores transition from proficient or highly effective to scores that indicate they are ineffective or emerging proficient. There are arguments to be made from both sides of the statistical aisle regarding whether the type of movement typically found in VAM data sets is acceptable or not. However, those types of discussions are largely academic in nature and isolated from the teachers who are subject to the conclusions derived via VAM. The American Statistical Association recognizes that ranking teachers by VAM scores can have negative consequences that reduce quality of instruction (ASA, 2014) and any legislative report would be expected to take that admonition into consideration.

Finding 3: The correlations provided in the Report vary widely

The Report also provides a set of between-year correlations of "teacher effects" by content area. Those correlations varied widely, ranging from a low of .35 to a high of .64, indicating that the range of explained variance in between-years "teacher effects" fell somewhere between 12.5% and 40.2%. The Report indicates these correlations demonstrate "moderate stability" (LDOE, 2019, p. 13). That said, the interpretation of moderate stability is not provided in the report. This omission requires any conclusion regarding stability to be subjective. Cohen (1988), for example, classified correlations between .3 and .5 as moderate, while Ratner (2009) suggested a moderate correlation is obtained with values between .3 and .7. Other sources suggested that correlation coefficients of .7 or lower do not demonstrate reliability and validity (Post, 2016).

As the literature documents, the notion of moderate stability does not enjoy a consensus position and the audience for the report would be better served by acknowledging that fact. It should be noted that the Report never defines "teacher effects," but it uses correlations with them to justify the validity of the model. This should be clearly defined for readers since VAM models, and consequently the calculation and interpretation of teacher effects, vary widely from state to state and model to model. Regardless of the chosen definition of "moderate stability," we have qualms with considering "moderately stable" a satisfactory benchmark for a high-stakes teacher evaluation algorithm, unless further explanation of what moderate stability is determined to be is forthcoming.

Finding 4: There was one data pool for teachers regardless of their teaching context

According to the Report, a database was compiled "based on a longitudinal data set derived from all students who took state-mandated tests in grades 4 through 10" (LDOE, 2019, p. 5). The VAM ranking assigned to a teacher was determined by grouping all teachers into this dataset, regardless of their grade level and content area. The pooling of all VAM-evaluated teachers into a centralized database is problematic. Haertel (2013) explained: Comparisons should be limited to fairly homogeneous groups of teachers. Rankings that mix teachers from different grade levels or teaching in schools with very different demographics places severe demands on statistical model assumptions, and the effects of violations of these assumptions are often not well understood. Conservative, local comparisons restricted to single subject areas and grade levels within homogeneous districts are much safer. (p. 25)

The recommendation from Haertel is methodologically sound and intuitive. Notwithstanding these facts, the Report indicates that data in Louisiana are indeed compiled into one broad dataset. While the Report does detail a variety of student and classroom context control variables that are included in the models, it does not address the biases of the origins described by Haertel.

Finding 5: Critically significant statistical information is missing

The most concerning finding was that multiple recommendations put forth by the ASA (2014) regarding VAM use were not followed in the Report. There were no reports of confidence intervals or standard errors, nor were there any reports on the amount of variance explained by the models or other estimates of model fit. Because of these omissions, there is no way to ascertain the precision of the VAM models used in Louisiana from the Report. There was also no discussion of potential sources of biases in the models. Information was presented about the variables entered into the models and coefficients were provided for the fixed effects in the hierarchical linear models used, but there was no reporting of statistical significance for the predictors. The LDOE presents the VAM models in Louisiana as valid measures of teacher effectiveness but fails to provide proper statistical justification to make that assertion.

Conclusions

Currently, the implementation of VAM in Louisiana is moving forward without requisite evidence for the effort's efficacy. The program that derives VAM outcomes is treated as proprietary and the results of the analyses were not made available to us through our public records request. No explanation for why the information was not shared was provided, even though this is an expectation of Louisiana's public record law (Louisiana Public Records Law, 2019, §44:32[D]).

The Report showed that the model is fluid and moves teachers around from category to category with some regularity. It also showed that the correlations for teachers from year-to-year significantly varied by subject matter, posing additional concerns about the fairness of VAM for all teachers. Furthermore, since we do not have sufficient information about the model, we are left with the myriad concerns related to model specification and fit that are prevalent in the research on VAM (e.g., Amrein-Beardsley, 2008; Goldhaber et al., 2013; Hill et al., 2011; Kersting et al., 2013; Schochet& Chiang, 2011).

Students are not "clean slates." We know there is a residual effect from prior teachers on student achievement that accounts for somewhere between 1% to 14% of the variability in test scores (ASA, 2014). Chetty et al. (2011) estimated 30% of teacher effect persists for up to three to four years, and that effect tends to remain somewhat constant for the remainder of a student's schooling. Yeh (2013) highlighted the additional obstacle of trying to dissociate school effects from teacher effects and concluded the dissociation of school and teacher level effects on student learning does not appear to be possible, absent a teacher's assignment to multiple schools with consistent evaluation systems. Very few teachers subject to VAM would satisfy the stipulation.

We note that nationally there appears to be a flux in the use of—and sentiments towards—VAM. Alaska, Arkansas, Kansas, Kentucky, North Carolina, and Oklahoma have recently stopped requiring student-growth measures in formal evaluations. Connecticut, Nevada, and Utah have adopted a modified response, requiring some evidence of student learning while disallowing the use of standardized-test scores to create that evidence (Loewus, 2017). In the particular case of Louisiana and its use of VAM, there are grounded concerns regarding the weight of decisions based on VAM and the lack of transparency regarding the process that calculated the results.

The legislature, as well as the citizens of Louisiana, deserve clear reporting that answers the concerns previously noted in our findings. In particular, the Report should disclose how much variance is being explained in these VAM efforts and how much range in that variance can be found across the spectrum of teachers. All other findings are subsumed by this information. It does not matter how many students took tests, how many teachers are part of the data pool, or even whether the scores are stable from year-to-year if the variance explained is not sufficient to support the idea that 35% of teacher evaluations in Louisiana should be informed by VAM results. Similarly, it appeared as though the Report is intended to suffice for public records requests from researchers as well. If this is the case, then it is imperative that the LDOE writes the Report for a variety of consumers—one that is comprehensible to someone with little-to-no statistical training and expertise while simultaneously incorporating all the technical information that is expected in statistical reporting for trained researchers. The ASA (2014) guidelines specifically detail best practices for these types of reports.

The questions raised and the issues identified in this paper appear to be germane for any state or school district employing VAM as an identifier of teacher quality; therefore, these issues need to be considered by policymakers and politicians in Louisiana and elsewhere. We also believe this paper provides an important prototype for other researchers to follow suit and examine the underlying assumptions and conclusions provided by the particular VAM efforts they analyze in other settings.

References

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Retrieved fromhttps://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Arsement, G. (2018, April 16). Blowing the whistle on HB-343. *Educate Louisiana.* http://educatelouisiana.org/2018/04/16/blowing-the-whistle-on-hb-343/

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65-75.

Amrein-Beardsley, A. (2018, February 22). New Mexico's motion for summary judgement, following Houston's precedent-setting ruling. Retrieved fromhttp://vamboozled.com/new-mexicos-motion-for-summary-judgment-following-houstons-precedent-setting-ruling/

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Educational Testing Service.

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: The National Academies Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (NBER Working Paper No. 17699). Cambridge, MA: The National Bureau of Economic Research.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220-236.

Goldhaber, D., & Hansen, M. (2008). Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions (Working Paper 31). National Center for Analysis of Longitudinal Data in Education Research.

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student scores. Educational Testing Service.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Houston Federation of Teachers, Local 2415, et al. v. Houston Independent School District, 251 F. Supp.3d 1168 (2017).

Kersting, N. B., Chen, M., & Stigler, J. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7), 1-39.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? *Education Finance and Policy, 6*(1), 18-42.

Loewus, L. (2017, November 14). Are states changing course on teacher evaluation? Test score growth plays lesser role in six states. *Education Weekly, 37*(13), 16-17. https://www.edweek.org/ew/articles/2017/11/15/are-states-changing-course-on-teacher-evaluation.html

Louisiana Public Records Law, R.S. 44 § 32 (2019).

Louisiana Department of Education. (2019, March 1). Louisiana's value-added assessment model for educator evaluation and support: A report in response to R.S. 17:3883(A)(8). Retrieved from https://www.louisianabelieves.com/docs/default-source/teaching/value-added-report-february-2019-final.pdf?sfvrsn=91be9f1f_4

Louisiana Department of Education. (n.d.). Value-added model. Retrieved June 27, 2019, from https://www.louisianabelieves.com/academics/value-added-model

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Obama White House Archives. (n.d.) Race to the top. https://obamawhitehouse.archives.gov/issues/education/k-12/race-to-the-top

Puffer, N. A., &Am rein-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.

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Post, M. W. (2016). What to do with "moderate" reliability and validity coefficients? *Archives of Physical Medicine and Rehabilitation, 97*(7), 1051-1052.

Ratner, B. (2009). The correlation coefficient: Its values range between +1/-1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing, 17*(2), 139-142.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* University of Tennessee Value-Added Research and Assessment Center.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*(2), 142-171.

Sentell, W. (2017, August 5). Louisiana teachers to face tougher job reviews in new school year under controversial evaluations. *The Advocate.*https://www.theadvocate.com/baton_rouge/news/politics/article_b80af760-76d5-11e7-b99a-9b931f3ecda5.html

U.S. Department of Education. (2009, November). *Race to the Top program executive summary*. https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Yeh, S. S. (2013). A reanalysis of the effects of teacher replacement using value added-modeling. *Teachers College Record, 115,* 1-35.