eJEP
eJournal of Education Policy

The Potential Consequence of Using Value-Added Models to Evaluate Teachers

Zuchao Shen
University of Cincinnati
PO Box 210002
Cincinnati, Ohio 45221
zuchao.shen@gmail.com

Carlee Escue Simon
University of Cincinnati
PO Box 210002
Cincinnati, Ohio 45221
carlee.escue@uc.edu

Ben Kelcey
University of Cincinnati
PO Box 210002
Cincinnati, Ohio 45221
benjamin.kelcey@uc.edu

Fall 2016

Abstract

Value-added models try to separate the contribution of individual teachers or schools to students' learning growth measured by standardized test scores. There is a policy trend to use value-added modeling to evaluate teachers because of its face validity and superficial objectiveness. This article investigates the potential long term consequences of making high-stakes decisions based on value-added teacher evaluations. To investigate this question, we analyze the micro-level effects on teacher effectiveness from the view of policy implementation and the macro-level effects on teacher quality based on the dynamic change of the teacher job market. We argue that the establishment of a formal connection between value-added teacher evaluations and high-stakes decision-making may compromise teacher effectiveness and teacher quality. We conclude that this connection between value-added measures and high-stakes decisions should not be established when it compromises the perception of a teacher's position as a secure and decent job.

Keywords: value-added, teacher evaluation, teacher effectiveness, teacher quality

Introduction

There are increasing interests among researchers and policymakers in using student growth data measured in standardized tests to evaluate the contribution and effectiveness of teachers and schools (Ballou & Springer, 2015; Harris & Herrington, 2015). The current shift from student-level accountability to teacher-level accountability and the focus of the "value" teachers "add" to student learning over time are largely driven by education reform advocates and the former U.S. Secretary of Education, Arne Duncan, and his Race to the Top competition (Collins & Amrein-Beardsley, 2014). The District of Columbia, over 40 states, and Puerto

Rico had applied and renewed waivers to not meet the prior goal of No Child Left Behind that all students would be academically proficient by the year 2014 in exchange for rigorous state-developed plans with stronger accountability mechanisms, these states are (or will be) using student growth scores as one component of their new teacher evaluation system (Collins & Amrein-Beardsley, 2014; U.S. Department of Education, 2016).

Researchers and policymakers have asserted the use of value-added scores to inform high-stakes decision-making, such as merit pay, tenure, promotion, and dismissal (Goldhaber, 2015). Value-added models (VAMs) are statistical models that analyze students' scores on standardized tests to separate the contributions to student growth made by individual teachers or schools from other factors beyond the control of teachers or schools (Darling-Hammond, 2015). In order to improve teacher effectiveness and teacher quality, which is believed to be beneficial for improving students' achievement, policymakers propose the use of value-added measures to inform high stakes decisions. Some school districts are using value-added measures to inform high-stakes decisions in teacher resource management, such as dismissal, retention, tenure or compensation. In some states and districts, student growth or value-added measures equal up to 50 percent in the teacher evaluation system (e.g., Florida Department of Education, 2015; Ohio Department of Education, 2015).

Policymakers have suggested that individual teachers are one of the most important factors to improve student learning outcomes, and can be influenced directly by policy (US Department of Education, 2009). In the use of VAMs for teacher evaluations and the use of evaluation results to inform high-stakes decision-making, policymakers have asserted that value-added teacher evaluations can improve teacher effectiveness and teacher quality, which in turn will improve student learning outcomes. But, teachers are only one of these factors that contribute to student learning, and is far away from the most important one; only about 10% of the variation in student learning outcomes measured by standardized test scores is attribute to teachers (American Statistical Association, 2014; Nye, Konstantopoulos, & Hedges, 2004).

The reason researchers and policymakers are interested in using VAMs to evaluate the contribution and effectiveness of teachers and schools is that VAMs show face validity and superficial objectiveness (Goldhaber & Theobald, 2013; Goldring et al., 2015). Face validity means that the measurement looks like it is covering the concept it purports to measure. Without thoughtful investigation, value-added teacher evaluations seems to measure what it supposes to measure, which is teacher effectiveness on students' academic growth. Superficial objectiveness means that the process and results of VAMs seem to be objective. Unlike other traditional teacher evaluation methods (e.g., classroom observation, supervisor rating) that involve raters who may subjectively influence the evaluation scores directly, VAM scores of individual teachers or schools are identical after the formulas and tests are set up. However, the results of VAMs vary from one another when different formulas and tests are used (Goldring et al., 2015). The formula and test used for value-added teacher evaluations are subject to choice by policymaker and experts. So, value-added teacher evaluations are not as objective and valid as may seem (Goldring et al., 2015).

The implementation of policy in value-added teacher evaluations also leads to some strategic actions made by teachers or schools when teachers or schools try to improve their VAM scores directly rather than through improvement of student learning outcome measured by test scores (Ballou & Springer, 2015). When using value-added teacher evaluation scores to inform high-stakes decision-making, the key question is whether teacher effectiveness and teacher quality could be improved in the long run. We argue that the consequence of using VAM scores to make high-stakes decisions may compromise teacher effectiveness and teacher quality in the long run, and, ultimately, negatively impact student learning outcomes measured in standardized tests.

In this article, we first review the value-added teacher evaluations in practice. We then investigate the potential micro- and macro-level consequences of using VAM scores to make high-stakes decisions. For the micro-level analysis, we interpret and analyze individual teacher's perception, actions and reactions upon value-added teacher evaluation, and the effect on teacher effectiveness. For the macro-level analysis, we examine the potential effects on the teacher job market when high-stakes decisions are made based on value-added teacher evaluations. We end this article with suggestions that no use of value-added measure should inform high-stakes decisions and to avoid the use of value-added teacher measures only to inform high-stakes decisions that lead to incentive decisions such as tenure or compensation.

Value-added teacher evaluations as policy implemented in practice

The promising goals of collecting and using student growth data in standardized tests were to measure and improve the effectiveness of teachers and principals (US Department of Education, 2009), though studies have shown that teachers have some level of difference in impacting their students' learning outcomes measured by test scores (Kane & Staiger, 2008; Kupermintz, 2003; Nye, Konstantopoulos, & Hedges, 2004). However, the use of VAMs in teacher evaluations in which one component is students' standardized test scores is problematic among some teachers and principals and controversial among researchers (Ballou & Springer, 2015). Concerning the reliability and validity of measurement in VAMs, there are debates about the potential use of value-added teacher evaluations along with high-stakes decision-making regarding professional development, payment, recruitment, promotion, and dismissal (Harris, 2013; Harris & Herrington, 2015).

Studies have shown that the ranking of teachers based on student test score gains suffers from reliability and validity issues, and that ranking varies across different tests or data sources. For example, ranking of teachers will differ if the value-added teacher evaluation uses different tests (Loeb & Candelaria, 2012; Papay, 2011) and different models (Goldhaber, 2015; Goldhaber & Theobald, 2013). Also, results of value-added teacher evaluations varied across different test subjects, years, and student groups (Goldhaber, 2015; Loeb & Candelaria, 2012). Some understated assumptions of using VAMs are apparent, including, but not limited to, teacher homogeneity (Condie, Lefgren, & Sims, 2014) and controlling for all other factors that influence students' learning outcomes in the model (McCaffrey, 2012; Raudenbush, 2013; 2015). Because of these reliability and validity issues developed by using value-added teacher evaluations, American Federation of Teachers President Randi Weingarten has called for the end of using value-added measures as a component of the teacher evaluation system (Sawchuk, 2014; Ballou & Springer, 2015).

It is not legitimate to criticize the use of VAMs in teacher evaluations merely based on reliability and validity concerns because other teacher evaluation methods also have similar reliability and validity problems (Harris, 2013). For example, that teachers get different scores when using different tests in VAMs is similar with teachers receiving different scores from different raters in classroom observations (Kelcey, McGinn, & Hill, 2014). The question is not whether we should use VAMs or not, the key question is how to use VAMs properly in teacher evaluations at the teacher and school levels, specifically in improving teacher effectiveness and teacher quality. What are the problems in practice when using VAMs in teacher evaluation? How can we implement VAMs in teacher evaluations properly in order to improve teacher effectiveness and teacher quality? What changes can we make to properly implement value-added teacher evaluations?

Value-added models and teacher effectiveness at the micro-level

Previous studies have shown that the consequences of policy initiatives, even the most promising ones, depend largely on what happens as individuals in the policy system interpret and act upon these initiatives (Goldhaber, 2015; Jiang, Sporte, & Luppescu, 2015; McLaughlin, 1987). When policy is implemented, it evolves through individuals' actions and reactions based upon the perception of what changes are required for them, what options are available, and what is the best choice for action (Spillane, Reiser, & Reimer, 2002).

For micro-level analysis, we investigated the potential consequence of using VAM results to make high-stakes decisions by analyzing individual teachers' perceptions, actions and reactions. We defined teacher effectiveness as how well in-service teachers delivered curriculum to assist students in growing mentally and intellectually as measured by student learning objectives (Goldhaber, 2002). Teacher effectiveness is largely influenced by teacher quality (e.g., the quality of applicants, the quality of the current teacher), pre- and in-service training opportunities, perception, and motivation and action in teaching (Goldhaber, 2015). When using VAMs to evaluate teachers, can value-added teacher evaluations, along with high-stakes decision-making, improve teacher effectiveness? In order to answer this question, we investigated teachers' perceptions of VAMs and their correspondent actions and reactions.

Teachers' perceptions of VAMs

Teachers' perceptions of VAMs are important because their actions and reactions toward VAMs are based upon what changes are required of them and what options must be acted upon. In a study of teachers' perceptions on a reform to evaluate teacher performance in Chicago (Jiang, Sporte, & Luppescu, 2015), 75% of the teachers linked with positive perception about the overall reform were accounted by observations on professional practice. However, teachers were skeptical about the inclusion of student growth data in the evaluation even when the proportion of value-added scores counted for the whole score was set from 15% to 0% dependent on which category teachers were evaluated. Teachers were concerned about the narrow representation of student growth that was measured by standardized tests and the increase in the already heavy testing burden placed upon them and their students (Jiang, Sporte, & Luppescu, 2015).

Teacher perceptions are influenced by the format of data dissemination (Jacobsen, Snyder, & Saultz, 2014) and the lack of transparency of the analytic engines that produce VAM scores arouse suspicion amongst teachers (Goldring et al., 2015; Jiang, Sporte, & Luppescu, 2015). Teachers' perceptions of evaluations are often related to and influenced by the perceptions of their principal and professional community (Jiang, Sporte, & Luppescu, 2015). In a study by Goldring et al. (2015), it was found that principals were most concerned about the VAM scores' timing (e.g., scores are not available when most employment/human resources decisions are made), validity and utility (e.g., whether VAM scores can measure teacher effectiveness, and whether the VAMs contained useful information for teachers to improve their instruction). Overall, principals were generally skeptical about VAM scores and were more likely to rely on observational results.

When value-added teacher evaluations are linked with high-stakes decision-making with direct-incentive effects, such as promotion, tenure, and compensation, it is quite possible that teachers with effective scores will welcome value-added teacher evaluation because of the financial incentive and extra monetary gains from it. However, teachers' perceptions will shift if the connection between VAM scores and potential benefits-cutting was set up at the individual teacher level. For example, teachers feel more comfortable and secure about their jobs if the scores are not used for direct-disincentive employment decisions (e.g., dismissal, payment, or punishment). If teachers with lower scores in VAM evaluations are supported by teachers with higher scores, the acceptance of value-added teacher evaluations is increased among teachers. However, if the scores in VAMs are used for direct-disincentive employment decisions (e.g., dismissal, payment, or punishment), the misclassification in using VAMs convinces teachers, even those who are effective teachers, to hold back their support because their job security is uncertain.

Teachers' actions, reactions and potential consequences

With a negative perception among teachers of the value-added model of teacher evaluations, it is not surprising that some teachers are potentially gaming the evaluation system to achieve high scores in VAMs. For example, a study by Ballou & Springer (2015) found that teachers coached students during tests when monitoring their own students in the exam. In the roster verification process, teachers excluded low-score students whom they had taught or included high-score students whom they had never taught to boost their scores. Teachers increased their scores in VAMs by using these strategies in ways that were counter to what was expected by policymakers (i.e., improving teaching effectiveness to have students learn more and perform better on tests). It is assumed that, when using VAM scores to make high-stakes decisions, teachers will be motived to improve students' learning if they received monetary rewards for students' test score gains. However, there is little to no evidence that this would be true (Baker, 2010).

Generally, standardized tests using item response theory can accurately measure students' capabilities, and these target students are usually in the middle part of the student population. Well-designed standardized differentiate the difference in the normal range of capabilities. However, standardized tests may do a poor job in differentiating the difference among lower bottom or upper top students. A study by Darling-Hammond (2015) showed that the use of value-added teacher evaluations also discouraged teachers who taught high-need students who had lower levels of achievement gains and who were always in the bottom of these tests, and teachers who taught gifted students whose achievement gains could hardly be measured in standardized test scores when they were always at the top of these tests.

The use of value-added teacher evaluations to make high-stakes decisions has the potential to improve in-service teacher effectiveness through direct-incentive or direct-disincentive decisions. Teachers may potentially improve their teaching effectiveness when they know their scores and can use available resources to improve their instruction. However, another concern arises that these incentive initiatives are linked with VAM measurements and may compromise the connection between teachers and their work environment, which in turn makes the work environment unpleasant and, ultimately, will damage school quality (Johnson, 2015). It is hard to conclude that the emphasis on individual-level improvements will lead to school-wide improvements in teaching effectiveness. It is quite possible that implementing VAM in teacher-level evaluations would actually lead to the decrease of teaching effectiveness school-wide, even though some individual teachers may dramatically improve their teaching quality (Johnson, 2015).

Although value-added teacher evaluations may be a useful tool to diagnose teachers' effectiveness, teachers' actions and reactions did not follow the expectation of how VAM policy would work. Our analysis at the micro-level indicated that the use of value-added teacher evaluations to inform high-stakes decisions may compromise the cooperative environment school-wide and may have no effect on improving teachers' effectiveness. The micro-level consequence is driven by individuals' perceptions, actions, and reactions in the pool of options or choices among applicants, pre-service teachers, in-service teachers, and school leaders.

Value-added model improvement on teacher quality at macro-level

For macro-level analysis, we examined the long-term effects of using value-added teacher evaluations to inform high-stakes decisions on the teacher job market. The macro-level consequence is driven by dynamic change in the perception of, and interaction between, a teaching career and other occupations. Researchers found a clear trend that the overall quality of the teaching workforce had constantly decreased over decades as compensation differed dramatically from the private sector (Goldhaber, 2015). For example, students who major in education tend to be drawn from the lower end of the ability distribution in standardized test scores (e.g., SAT and ACT). The average test scores of those who enter teaching are almost 10% lower than that of those entering other professions. College graduates with high test scores are less likely to pursue a teaching career (Goldhaber, 2002).

To improve the overall quality of the teacher workforce, a simulation study found that student learning outcomes measured by standardized scores would have increased dramatically if 5% to 10% if the least effective teachers were dismissed and replaced annually with teachers who are of average effectiveness (Hanushek, 2009). Other simulation studies drew similar conclusions on students' earnings (Goldhaber & Hansen, 2010; Chetty, Friedman, & Rockoff, 2014). When it comes to the use of value-added teacher evaluations for such high-stakes decision-making, even ignoring the misclassification by VAM that will incorrectly dismiss non-tenured teachers, Ronfeldt, Loeb, & Wyckoff (2012) show that higher teacher turnover rates had a disruptive effect on student achievement.

Furthermore, the position of teachers has historically been viewed as a very predictable and decent occupation measured by payment and job security. The use of VAM scores to dismiss teachers may compromise this perception. The understanding of high turnover rates associated with teaching positions and low pay may make this career less attractive to new applicants. In turn, this decreases overall teacher quality if the average compensation of teachers has not been raised dramatically (Goldhaber, 2015).

The solution to improve teacher quality is to compete with the private sector for new and smarter applicants, along with a selective procedure in the process of recruitment. In-service training is also important, but in-service training exists in almost every decent job position. Extra in-service training with a more advanced training than other sectors may be a second choice to improve teacher quality, but it may have less effect in attracting more qualified applicants. The key to improving teacher quality is in compensation packages and in the perception of teaching as a decent secure job position. Without more investment in teacher salaries, improving teacher quality is like a dog chasing its own tail; teacher quality is determined by the market and investment in teachers at a macro-level.

In addition, how principals act may determine the ultimate direction of policy implementation at the school level. Principals may not act upon what policy expectations as some teachers may, however, principals' options are also largely dependent on available resources. Principals may not fire teachers if they see or predict the negative effects the dismissal on the morale of teachers or their perceptions of job security. Some principals strategically arrange teachers who showed effectiveness in VAM evaluation to teach testing subjects and teachers who have low VAM scores to teach untested subjects. Principals would prefer to incentivize teachers if they had the extra resources to do so, but would rather not do so if required to use the resources derived from cutting from the ineffective teachers.

Discussion

As we have mentioned before, policies evolve in the process of implementation, and it is very important to constantly evaluate current educational policy on value-added assessments. The goals of program evaluations are to provide solid evidence to decide whether such a program should be modified, improved, or terminated (Windsor, 2015). All policies should be influenced or affected by evidence-based evaluation results so policies can be implemented to fulfill proposed goals and objectives. All states should evaluate their value-added models or student growth measures in their teacher evaluation systems to make sure the most robust models are used, reflecting the contribution of teachers and that all shareholders have been trained with knowledge and understanding about the mechanism of value-added measures and tools to use in the evaluation systems.

It also should be mentioned that value-added evaluations should be based on high-quality standardized tests to measure student growth between two points in time; no state has developed state tests that can cover all grades and subjects so far. For subjects and grades not covered by state tests, some states use alternative tests (e.g., tests from testing organizations or tests developed by districts) to test students and use these results to evaluate teachers. It is important to take steps to control the quality of alternative tests so that the reliability, validity and fairness of test can be ensured.

As most states are using or piloting value-added or student growth measures as one component of new teacher evaluation systems, there are several versions of value-added models used by different states (e.g., student growth percentiles model, the SAS education value-added assessment system [SAS EVAAS], and the Value-Added Research Center [VARC] model, (see Collins & Amrein-Beardsley, 2014)). Since some states use student-expected percentiles to measure growth without controlling for any covariates, the measured percentile difference in tests between two points in time actually reflect overall effect by all factors that influenced student achievement (e.g., small class size versus big class size or the change of school investment in classroom). It is not appropriate to use these overall effects to evaluate teachers. Future research on pros and cons of different value-added models will definitely benefit the field as it will provide states with accessible knowledge to improve their value-added models in practice.

Conclusion

As Goldhaber (2002) pointed out "good teachers certainly make a difference, but it's unclear what makes for a good teacher" (p2). Through our analysis, we concluded that using value-added teacher evaluations to inform high-stakes decision-making may not make for a good teacher. Without extra resources to invest in it, VAM itself may not solve the problem that policymakers intend to solve. Even with extra resources invested in education, implementing new policy without attending to the consequences at the micro- and macro-levels, expected results could be disastrous. As suggested by Jiang, Sporte and Luppescu (2015), the clarity and practicality (i.e., instrumentality, congruence, and cost) in policy implementation are the key factors that drive the direction of an individual's perception. These perceptions may direct their actions and reactions against what policymakers expect.

The implementation of new policy without fully considering consequences in micro- and macro-levels may result in the opposite consequences to what is anticipated becoming reality. Micro-level consequences are driven by an individual's perceptions, actions, and reactions when regarding a pool of applicants, pre-service teachers, in-service teachers, and school leaders. Macro-level consequences are driven by a dynamic change in the perception of, and interaction between, a teaching career and other occupations.

High-rate negative high-stakes decision-making (e.g., dismal or low pay) would compromise the nature of teaching positions, which, in turn, may decrease teacher quality if there were no significant benefits or improvements associated with teaching positions to counterbalance the negative effects associated with such decision-making. Positive high-stakes decision-making (e.g., promotion or tenure) could possibly improve teacher quality individually and school-wide if more resources are available and policies are carefully designed and implemented. Therefore, unbalancing incentives/disincentives to be conditional with school-wide improvements would benefit individual teachers and the whole school. In conclusion, we call for caution when using value-added teacher evaluations to inform direct incentive high-stakes decision-making (e.g., tenure, promotion, and recruitment), and to not implement them to inform disincentive high-stakes decision-making (e.g., dismissal, payment, or punishment).

References

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Alexandria, VA: Author. *Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.*

Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, 45, 2, 123-137.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher*, 44(2), 77-86.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. *EPI Briefing Paper# 278. Economic Policy Institute.*

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633-2679.

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1), 1-34*.*

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40, 76-92.

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

Florida Department of Education. (2015). Performance Evaluation. Available at http://www.fldoe.org/teaching/performance-evaluation

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, 2(1), 50-55.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. Educational Researcher, 44(2), 87-95.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing?. *American Educational Research Journal*, 47(1), 218-251.

Goldhaber, D., & Theobald, R. (2013). Do different value-added models tell us the same things? Carnegie Knowledge Network Briefs. Stanford, CA. Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Goldhaber_Nov2013-Update.pdf

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, 44(2), 96-104.

Hanushek, E. A. (2009). Teacher deselection. In D.Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165-180). Washington, DC: Urban Institute Press.

Harris, D. N. (2012). How do value-added indicators compare to other measures of teacher effectiveness? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 5*. Carnegie Foundation for the Advancement of Teaching*.

Harris, D. N. (2013). How might we use multiple measures for teacher accountability? Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/11/CKN_2013_10_Harris.pdf

Harris, D. N., & Herrington, C. D. (2015). Editors’ introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher*, 44(2), 71-76.

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education*, 121(1), 1-27.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform Chicago's REACH students. *Educational Researcher*, 44(2), 105-116.

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school?. *Educational Researcher*, 44(2), 117-126.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working paper14607). Cambridge, CA: National Bureau of Economic Research. Available at http://public2-prod.gsb.stanford.edu/sites/default/files/documents/ame_05_09_staiger.pdf

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, 5.

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement*, 13, 3, 291-325.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis*, 25, 3, 287-298.

Loeb, S. (2013) How can value-added measures be used for teacher improvement? (Carnegie Knowledge Network, what we know series). Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/12/CKN-Loeb_Teacher-Improvement.pdf

Loeb, S., & Candelaria, C. (2012). How stable are value-added estimates across years, subjects, and student groups?. Carnegie Knowledge Network. Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Loeb.pdf

McCaffrey, D. F. (2012). Do value-added methods level the playing field for teachers? Carnegie Knowledge Network. Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/06/CKN_2012-10_McCaffrey.pdf

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational evaluation and policy analysis*, 9(2), 171-178.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational Evaluation and Policy Analysis*, 26, 3, 237-257.

OECD. (2014). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science*, PISA, OECD Publishing. Available at http://dx.doi.org/10.1787/9789264208780-en.

Ohio Department of Education. (2015). *Value-added student growth measure*. Available at http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Student-Growth-Measures/Value-Added-Student-Growth-Measure

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

Raudenbush, S. W. (2013). What do we know about using value-added to compare teachers who work in different schools? Available at http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/08/CKN_Raudenbush-Comparing-Teachers_FINAL_08-19-13.pdf

Raudenbush, S. W. (2015). Value added a case study in the mismatch between education research and policy. *Educational Researcher*, 44(2), 138-141.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, 50(1), 4-36.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of educational research*, 72(3), 387-431.

Sawchuk, S. (2014). AFT's Weingarten backtracks on using value-added measures for evaluations. *Education Week*. Available at https://blogs.edweek.org/edweek/teacherbeat/2014/01/weingartens_retrenchment_on_va.html

US Department of Education. (2009). Race to the top program: Executive summary. Available at https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2016). ESEA flexibility [Web page]. Washington DC. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Windsor, R. (2015). *Evaluation of health promotion and disease prevention programs: Improving population health through evidence-based practice (5th ed.)*. New York, NY: Oxford University Press.