Brigham Young University Education and Law Journal

Volume 2012 | Issue 1 Article 2

Spring 3-1-2012

The Legal and Policy Implication of Value-Added Teacher Assessment Policies

Preston C. Green

Bruce D. Baker

Joseph Oluwole

THE LEGAL AND POLICY IMPLICATIONS OF VALUE-ADDED TEACHER ASSESSMENT POLICIES

*Preston C. Green III**

*Bruce D. Baker***

*Joseph Oluwole****

I. INTRODUCTION

In the past few years, numerous political think-tanks have claimed that teacher evaluation systems must be strengthened to prevent the granting of tenure to incompetent teachers.[1] Several states have responded to these criticisms by requiring teachers to be evaluated in part based on the academic achievement of their students.[2] Colorado, Louisiana, and Tennessee have required their teacher evaluation systems to be

* Harry Lawrence Batschelet II Chair Professor of Educational Administration, Professor of Education and Law, Penn State University.

** Associate Professor of Education, Rutgers University.

*** Assistant Professor of Education, Montclair State University.

1. Authors of The Widget Effect from The New Teacher Project, in a study of twelve districts in four states, claim that 99% of tenured teachers in districts using a satisfactory/unsatisfactory evaluation system received a positive rating. The same study claims that in districts with more ratings options, 94% of teachers still received the two highest rating options and less than 1% received a rating of unsatisfactory. DANIEL WEISBERG ET AL., THE NEW TCHR. PROJECT, THE WIDGET EFFECT: OUR NATIONAL FAILURE TO ACKNOWLEDGE AND ACT ON DIFFERENCES IN TEACHER EFFECTIVENESS (2d ed. 2009), *available at* http://widgeteffect.org/downloads/TheWidgetEffect.pdf. While the findings are striking, the study has come under fire for poor documentation of methods, leading to concerns that the findings are significantly overstated. RAYMOND L. PECHEONE & RUTH C. WEI, REVIEW OF "THE WIDGET EFFECT: OUR NATIONAL FAILURE TO ACKNOWLEDGE AND ACT ON TEACHER DIFFERENCES" (2009), *available at* http://nepc.colorado.edu/files/TTR-Pecheone-WIDGET.pdf.

2. This trend may be the result of President Obama's "Race to the Top Program," by which the Department of Education provides $4.35 billion for educationally innovative programming. It has also encouraged states to revamp their teacher evaluation policies to include student achievement data. Andrew J. Rotherham, *Rating Teachers: The Trouble with Value-Added Data*, TIME, Sept. 23, 2010, *available at* http://www.time.com/time/nation/article/0,8599,2020867,00.html.

based 50% or more on the academic growth of the students.[3] At the time of submission of this article, only a handful of states had taken such bold steps. But, by the time of our final edits, 18 states had overhauled teacher evaluation requirements with 14 requiring that at least 40% of teacher evaluation consist of student performance measures.[4] While the attempts to link teacher evaluations to student achievement may seem reasonable on their face, the primary approach for doing so, called value-added modeling (VAM), suffers from substantial technical problems that may result in an alarming number of good teachers being falsely identified as "ineffective" and eventually terminated. This may be especially true in states adopting policies requiring 50% or more of their teacher evaluations to be based on student achievement scores.

This article examines the framework and potential legal problems of such teacher evaluation policies. The second section below provides an overview of those states that have adopted evaluation programs that are based 50% or more on student achievement scores. The third and fourth sections identify some of the technical problems associated with value-added measures and discuss how teacher evaluation plans that overly rely on student achievement data may be vulnerable to legal challenges. The final section observes that random teacher assignments, though problematic for removing discretion from school districts, are the best way to minimize legal and other challenges to teacher evaluation policies that rely 50% or more on student achievement scores.

II. PROGRAMS IN WHICH STUDENT PERFORMANCE IS THE BASIS OF 50% OR MORE OF TEACHER EVALUATIONS

Colorado, Louisiana, and Tennessee have enacted teacher evaluation systems requiring 50% or more of the evaluations to be based on students' academic growth. This section summarizes the evaluation systems in these states, as well as the procedural protections provided for teachers deemed ineffective.

---

3. *See infra* Part II.

4. NAT'L COUNCIL ON TCHR. QUALITY, STATE OF THE STATES: TRENDS AND EARLY LESSONS ON TEACHER EVALUATION AND EFFECTIVENESS POLICIES (Oct. 2011), *available at* http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf.

A. Colorado

Colorado's statute on licensed personnel evaluations in the area of education creates a state council for educator effectiveness to advise the State Board of Education.[5] A major goal of this council is to aid in the creation of teacher evaluation systems that "will ensure that every teacher is evaluated using multiple fair, transparent, timely, rigorous, and valid methods."[6] Considerations of student academic growth must comprise at least 50% of each evaluation.[7] Quality measures for teachers must include "measures of student longitudinal academic growth," such as "interim assessment results or evidence of student work, provided that all are rigorous and comparable across classrooms and aligned with state model content standards and performance standards."[8] These quality standards must take into account diverse factors, including "special education, student mobility, and classrooms with a student population in which ninety-five percent meet the definition of high-risk student."[9]

Colorado's statute also calls for school districts to develop appeals procedures for teachers and principals found wanting.[10] A teacher or principal who is deemed ineffective must receive written notice, the documentation used for making this determination, and identification of the deficiency.[11] Furthermore, the school district must ensure that a tenured teacher who disagrees with this designation has "an opportunity to appeal that rating, in accordance with a fair and transparent process developed, where applicable, through collective bargaining."[12] If no collective bargaining agreement is in place, then the teacher may request a review "by a

---

5. COLO. REV. STAT. § 22-9-105.5(2)(a) (2010).

6. *Id.* § 22-9-105.5(3)(a).

7. *Id.*

8. *Id.*

9. *Id.* The statute also calls for the creation of performance evaluation councils to advise school districts. *Id.* § 22-9-107(1). The performance evaluation councils also help school districts develop teacher evaluation systems that must be based on the same measures as those developed by the state council for educator effectiveness. *Id.* § 22-9-106(1)(e)(II). However, the performance evaluation councils lose their authority to set standards once the state board has promulgated rules and the initial phase of statewide implementation has been completed. *Id.* § 22-9-106(1)(e)(I).

10. *Id.* § 22-9-106(3.5)(b)(II).

11. *Id.* § 22-9-106(3.5)(b)(I).

12. *Id.* § 22-9-106(3.5)(b)(II).

mutually agreed-upon third party."[13] The school district or Board for Cooperative Services must develop a remediation plan to correct these deficiencies to include professional development opportunities that are intended to help the teacher achieve an effective rating in her next evaluation.[14] The teacher or principal must receive a reasonable amount of time to correct such deficiencies.[15]

B. Louisiana

Louisiana's Professional Employee Quality Development Act requires every teacher to be evaluated annually by a local school board.[16] By the start of the 2012-2013 school year, 50% of teacher evaluations will be based on evidence of student-achievement growth "using a value-added assessment model as determined by the board for grade levels and subjects for which value-added data is [sic] available."[17] Where value-added data are unavailable, the board will establish the growth measures. The model must take into account other factors, including students with disabilities, students eligible for free and reduced lunch, student attendance, and student discipline.[18]

The board must place teachers who are found to be ineffective in an intensive assistance program[19] after informing teachers in writing of their need for intervention.[20] The assistance program must include at a minimum: (1) steps needed for the teacher to improve; (2) assistance that the board will provide; (3) a time line, not exceeding two years, to achieve the objectives; and (4) actions to be taken if the teacher fails to improve.[21] If the teacher fails to complete the assistance program in compliance with the Act, or if the teacher is deemed "ineffective after a formal evaluation conducted immediately upon completion of the program," then the board must "timely initiate termination proceedings."[22]

---

13. *Id.*
14. *Id.* § 22-9-106(3.5)(b)(I)-(II).
15. *Id.*
16. LA. REV. STAT. ANN. § 17:3902(A) (2010).
17. *Id.* § 17:3902(B)(5).
18. *Id.*
19. *Id.* § 17:3902(C)(2).
20. *Id.* § 17:3902(C)(2)(a).
21. *Id.* § 17:3902(C)(2)(b).
22. *Id.* § 17:3902(C)(2)(b)(v).

C. *Tennessee*

Tennessee's evaluation statute, titled "Tennessee First to the Top Act of 2010," creates a teacher evaluation advisory committee to develop an annual evaluation for all teachers.[23] The Act stipulates that 50% of the evaluation criteria must consist of student achievement data. Thirty-five percent (35%) of that percentage must be based on the Tennessee Value-Added Assessment System (TVAAS) or a comparable test for student growth if no TVAAS data are available.[24] The remaining 15% must be mutually agreed upon by the evaluator and the teacher being evaluated.[25] Other mandatory criteria for teachers include: (1) review of previous evaluations; (2) personal conferences, including discussion of strengths, weaknesses, and remediation; and (3) classroom or position observation followed by a written evaluation.[26] The Act also requires the committee to develop a local-level grievance procedure, which enables teachers to challenge the accuracy of the data used to evaluate the teacher and compliance with the statute's evaluation policies.[27]

III. TECHNICAL PROBLEMS OF VALUE-ADDED MODELING[28]

As noted in the prior section, three states base 50% of their teacher evaluation systems on student achievement data. Most

23. *See generally* TENN. CODE ANN. § 49-1 (2010).

24. *Id.* § 49-1-302(d)(2)(A)(i).

25. *Id.* § 49-1-302(d)(2)(A)(ii).

26. *Id.* § 49-1-302(d)(2)(B).

27. *Id.* § 49-1-302(d)(2).

28. In this article, we address specifically value-added modeling, a statistical technique which attempts to attribute (with causal inference) student learning gains to teachers of record for those students. A handful of states including Colorado have adopted a method referred to as "student growth percentile" scores which are a descriptive measure used to characterize student achievement growth including average student achievement growth of classes of students. These measures are not intended for making inferences about teacher effectiveness. *See* Bruce D. Baker, *Take Your SGP and VAMit, Damn it!*, SCHOOL FINANCE 101 (Sept. 2, 2011), http://schoolfinance101.wordpress.com/2011/09/02/take-your-sgp-and-vamit-damn-it/; Damian W. Betebenner et al., *Student Grown Percentiles and Shoe Leather*, EDUC. NEWS COLO. (Sept. 13, 2011), http://www.ednewscolorado.org/2011/09/13/24400-student-growth-percentiles-and-shoe-leather. Yet, state officials have proposed that these measures be used for evaluating teacher effectiveness. See Bruce D. Baker, Piloting the Plane on Musical Instruments & Using SGPs to Evaluate Teachers, SCHOOL FINANCE 101 (Sept. 22, 2011), http://schoolfinance101.wordpress.com/2011/09/22/piloting-the-plane-on-musical-instruments-using-sgps-to-evaluate-teachers/.

of the VAM teacher ratings thereby generated attempt to predict the influence of the teacher on the student's end-of-year test score, given the student's prior test score and descriptive characteristics—for example, whether the student is poor, has a disability, or is limited in her English language proficiency.[29] These statistical controls are designed to account for the differences that teachers face in serving different student populations.

There are, however, many problems associated with using VAM to determine whether teachers are effective. Among these problems are the instability of teacher ratings, classification and model prediction error, unreliable results from different "standardized" tests, difficulties in isolating a single teacher's contribution to students' learning, the non-random assignment of students across teachers, schools, and districts, and the struggle for teachers to even receive VAM ratings. This section details how these problems undermine the effectiveness of using VAM teacher ratings to evaluate the effectiveness of teachers.

A. Instability of Teacher Ratings

The assumption in VAM for estimating teacher "effectiveness" is that if one uses data on enough students passing through a given teacher each year, one can generate a stable estimate of the teacher's contribution to the students'

29. Value-added ratings of teachers are generally not based on a simple subtraction of each student's fall test score from the following spring's test score for a specific subject. Such an approach would clearly disadvantage teachers who happen to serve less motivated groups of students or students with more difficult home lives and/or fewer family resources to support their academic progress throughout the year. It would be even more problematic to use the spring test score from the prior year as the baseline score for comparison with the spring test score of the current year to evaluate the current teacher because the teacher had little control over any learning gain or loss that may have occurred during the prior summer. Additionally, these gains and losses tend to be different for students of higher and lower socioeconomic status. *See* Karl L. Alexander et al., *Schools, Achievement, and Inequality: A Seasonal Perspective*, 23 EDUC. EVALUATION & POL'Y ANALYSIS 171 (2001). Recent findings from a study funded by the Bill and Melinda Gates Foundation confirm these "seasonal" effects: "The norm sample results imply that students improve their reading comprehension scores just as much (or more) between April and October as between October and April in the following grade. Scores may be rising as kids mature and get more practice outside of school." BILL & MELINDA GATES FOUND., LEARNING ABOUT TEACHING: INITIAL FINDINGS FROM THE MEASURES OF EFFECTIVE TEACHING PROJECT 8 (2010), *available at* http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

achievement gains.[30] This assumption is problematic because of the concept of inter-temporal instability: that is, the same teacher is highly likely to get a very different value-added rating from one year to the next. The year-to-year correlation for a teacher's value-added rating is only about 0.2 or 0.3—at best a very modest correlation. Sass also notes that:

> About one quarter to one third of the teachers in the bottom and top quintiles stay in the same quintile from one year to the next while roughly 10 to 15 percent of teachers move all the way from the bottom quintile to the top and an equal proportion fall from the top quintile to the lowest quintile in the next year.[31]

Furthermore, most of the change or difference in the teacher's value-added rating from one year to the next is unexplainable—by differences in observed student characteristics, peer characteristics, or school characteristics.[32]

Similarly, preliminary analyses from the Measures of Effective Teaching Project, funded by the Bill and Melinda Gates Foundation, found:

> When the between-section or between-year correlation in teacher value-added is below .5, the implication is that more than half of the observed variation is due to transitory effects rather than stable differences between teachers. That is the case for all of the measures of value-added we calculated.[33]

While some statistical corrections and multi-year analysis might help, it is hard to guarantee or even be reasonably sure that a teacher would not be dismissed simply as a function of unexplainable low performance for two or three years in a row.

B. Classification and Model Prediction Error

Another technical problem of VAM teacher evaluation systems is classification and/or model prediction error. In a study funded by the U.S. Department of Education, researchers at Mathematica Policy Research Institute carried out a series

---

30. TIM R. SASS, THE STABILITY OF VALUE-ADDED MEASURES OF TEACHER QUALITY AND IMPLICATIONS FOR TEACHER COMPENSATION POLICY (2008), *available at* http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf. *See also* Daniel F. McCaffrey et al., *The Intertemporal Variability of Teacher Effect Estimates*, 4 EDUC. FIN. & POL'Y 572 (2009).

31. SASS, *supra* note 30, at 2.

32. *Id.*

33. BILL & MELINDA GATES FOUND., *supra* note 29, at 19.

of statistical tests and reviews of existing studies to determine the identification "error" rates for ineffective teachers when using typical value-added modeling methods.[34] The report found:

> Type I and II error rates for comparing a teacher's performance to the average are likely to be about 25 percent with three years of data and 35 percent with one year of data. Corresponding error rates for overall false positive and negative errors are 10 and 20 percent, respectively.[35]

Type I error refers to the probability that based on a certain number of years of data, the model will find that a truly average teacher performed significantly worse than average.[36] Thus, there is about a 25% chance if using three years of data or a 35% chance if using one year of data that a teacher who is "average" would be identified as "significantly worse than average" and potentially be fired. Of particular concern is the likelihood that a "good teacher" is falsely identified as a "bad" teacher—in this case a "false positive" identification. According to the study, this occurs one in ten times given three years of data and two in ten times given only one year of data.[37]

C. Same Teachers, Different Tests, Different Results

Determining whether a teacher is effective may vary depending on the assessment used for a specific subject area and not whether that teacher is a generally effective teacher in that subject area. For example, Houston uses two standardized tests each year to measure student achievement: the state Texas Assessment of Knowledge and Skills (TAKS) and the national Stanford Achievement Test.[38] NYU Professor Sean Corcoran and colleagues used Houston Independent School District (HISD) data from each test to calculate separate value-added measures for fourth- and fifth-grade teachers.[39] The

---

34. PETER Z. SCHOCHET & HANLEY S. CHIANG, U.S. DEP'T EDUC., ERROR RATES IN MEASURING TEACHER AND SCHOOL PERFORMANCE BASED ON STUDENT TEST SCORE GAINS (July. 2010), *available at* http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf.

35. *Id.* at i.

36. *Id.* at 12.

37. *Id.* at i.

38. SEAN P. CORCORAN ET AL., ANNENBERG INSTITUTE FOR SCHOOL REFORM, 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 13 (2010), *available at* http://annenberginstitute.org/pdf/valueaddedreport.pdf.

39. *See Id.* at 17.

authors found that a teacher's value-added rating can vary considerably depending on which test is used.[40] Specifically:

> [A]mong those who ranked in the top category (5) on the TAKS reading test, more than 17 percent ranked among the lowest two categories on the Stanford test. Similarly, more than 15 percent of the lowest value-added teachers on the TAKS were in the highest two categories on the Stanford.[41]

Recent findings from the Bill and Melinda Gates' Foundation's Measures of Effective Teaching (MET) project were quite similar to those of Corcoran.[42] While the MET report authors downplayed low correlations between teacher ratings generated by different tests, independent reviewer, University of California at Berkeley economist Jesse Rothstein explained:

> The study finds that the measures are related, but only modestly. The report interprets this as support for the use of value-added as the basis for teacher evaluations. This conclusion is unsupported, as the data in fact indicate that a teachers' [sic] value-added for the state test is not strongly related to her effectiveness in a broader sense. Most notably, value-added for state assessments is correlated 0.5 or less with that for the alternative assessments, meaning that many teachers whose value-added for one test is low are in fact quite effective when judged by the other.[43]

Similar issues apply to tests on different scales: different possible ranges of scores or different statistical modification or treatment of raw scores—for example, whether student test scores are first converted into standardized scores relative to an average score or expressed on some other scale such as percentile rank (which is done in some cases but would generally be considered inappropriate). For instance, if a teacher is typically assigned higher performing students and the scaling of a test is such that it becomes very difficult for students with high starting scores to improve over time, that teacher will be at a disadvantage. However, another test of the

---

40. *Id.*

41. *Id.*

42. BILL & MELINDA GATES FOUND., WORKING WITH TEACHERS TO DEVELOP FAIR AND RELIABLE MEASURES OF EFFECTIVE TEACHING (June 2010), *available at* http://www.metproject.org/downloads/met-framing-paper.pdf.

43. JESSE ROTHSTEIN, REVIEW OF *LEARNING ABOUT TEACHING* 1 (Jan. 2011), *available at* http://nepc.colorado.edu/files/TTR-MET-Rothstein.pdf.

same content or another test with a different scaling of scores (so that smaller gains are adjusted to reflect the relative difficulty of achieving those gains) may produce an entirely different rating for that teacher.

D. Difficulty in Isolating Any One Teacher's Influence on Student Achievement

It is difficult, if not entirely infeasible, to isolate one specific teacher's contribution to students' learning, leading to situations where a teacher might be identified as a bad teacher simply because her colleagues are ineffective. This is called the spillover effect.[44] For students who have more than one teacher across subjects (and/or teaching aides/assistants), each teacher's value-added measures may be influenced by the other teachers serving the same students. Northwestern University Professor Kirabo Jackson and researcher Elias Bruegmann, for example, found in a study of North Carolina teachers that students perform better, on average, when their teachers have more effective colleagues.[45] University of Missouri Professor Cory Koedel found that reading achievement in high school is influenced by both English and math teachers.[46] These spillover effects mean that teachers assigned to weaker teams of teachers might be disadvantaged through no fault of their own.

E. Non-Random Assignment of Students Across Teachers, Schools, and Districts

The fact that teacher value-added ratings cannot fully be disentangled from patterns of student assignment across schools and districts leads to the likelihood that teachers serving larger shares of one population versus another are more likely to be identified as effective or ineffective through no fault of their own. This non-random assignment problem relates not to the error in the measurement of test scores, but to the complications of applying a statistical model to real-

44. Cory Koedel, *An Empirical Analysis of Teacher Spillover Effects in Secondary School*, 28 ECON. EDUC. REV. 682 (2009).

45. C. Kirabo Jackson & Elias Bruegmann, *Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers*, 1 AM. ECON. J.: APPLIED ECON. 85 (2009).

46. Koedel, *supra* note 44, at 691.

world conditions. The fairest comparisons between teachers would occur in a case where teachers were randomly assigned to comparable classrooms with comparable resources and where exactly the same number of students were randomly assigned to those teachers. Teachers would then have the same number of children with similar family backgrounds, prior performance, personal motivation, and other characteristics.

Such circumstances, however, are unrealistic. Students are not sorted randomly across schools, districts, or teachers within schools. Nor are teachers randomly assigned across school settings with equal resources. Instead, it is likely that one fourth-grade teacher in a school is assigned more difficult students year after year than another. This may occur by choice of that teacher, having a desire to try to help such students, or by other factors, such as the desire of a principal to make a teacher's work more difficult. While most value-added models contain some crude indicators of poverty status, language proficiency, and disability classification, few, if any, sufficiently mitigate the bias that occurs as a result of non-random student assignment. Bias stems from such apparently subtle forces as the influence of peers on one another and the inability of value-added models to sufficiently isolate the teacher effect from the peer effect, both of which occur in the classroom.[47] In fact, University of California, Berkeley Professor Jesse Rothstein notes that "[r]esults indicate that even the best feasible value added models may be substantially biased, with the magnitude of the bias depending on the amount of information available for use in classroom assignments."[48]

---

47. There exist at least two different approaches to control peer group composition. One approach involves constructing measures of the average entry level of performance for all other students in the class. Caroline M. Hoxby & Gretchen Weingarth, Malcolm Wiener Inequality & Social Policy Seminar Series, *Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects* (Mar. 20, 2006), *available at* http://www.hks.harvard.edu/inequality/Seminar/Papers/Hoxby06.pdf. Another involves constructing measures of the average racial and socioeconomic characteristics of classmates. Eric A. Hanushek & Steven G. Rivkin, *School Quality and the Black-White Achievement Gap* (Nat'l Bureau of Econ. Research, Working Paper No. 12651, 2006), *available at* http://faculty.smu.edu/millimet/classes/eco7321/papers/hanushek%20rivkin%2002.pdf.

48. Jesse Rothstein, *Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables* (Nat'l Bureau of Econ. Research, Working Paper No. 14666, 2009), *available at* http://www.nber.org/papers/ w14666.pdf?new_window=1. *See also* Jesse Rothstein, *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*, 125 Q.J. ECON. 175 (2010). Many advocates

Value-added modeling has more recently been at the center of public debate after the *Los Angeles Times* contracted RAND Corporation economist Richard Buddin to estimate value-added scores for Los Angeles teachers,[49] and the *Times* reporters then posted the names of individual teachers classified as effective or ineffective on their web site.[50] The fairly typical model used by Buddin produced technical documentation rich with evidence of the types of model bias described by Rothstein and others.[51] For example:

- Ninety-seven percent of children in the lowest performing schools and 55% in higher performing schools are poor;
- The number of gifted children in a class affects the teacher's value-added estimate positively; —the more gifted children, the higher the teacher's effectiveness rating;
- Black teachers have lower value-added scores for both English Language Arts and math than white teachers;
- Having more black students in a class is negatively, albeit minimally, associated with teacher's value-added scores;
- Asian teachers have higher value-added scores than white teachers for Math, with a positive association

---

of value-added approaches point to a piece by Thomas Kane and Douglas Staiger as downplaying Rothstein's concerns. Thomas J. Kane & Douglas O. Staiger, *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (Nat'l Bureau of Econ. Research, Working Paper No. 14607, 2008), *available at* http://www.nber.org/papers/w14607.pdf?new_window=1. However, with regard to the Kane and Staiger analysis, Eric Hanushek and Steven Rivkin explain: "the possible uniqueness of the sample and the limitations of the specification test suggest care in interpretation of the results." Eric A. Hanushek & Steven G. Rivkin, Presentation for the American Economic Association, *Generalizations about Using Value-Added Measures of Teacher Quality* 7 (Jan. 3-5, 2010), *available at* http://www.utdallas.edu/research/tsp-erc/pdf/jrnl_hanushek_rivkin_2010_teacher_quality.pdf.

49. Richard Buddin, *How Effective Are Los Angeles Elementary Teachers and Schools?* (MPRA Paper No. 27366, Aug. 31, 2010), *available at* http://mpra.ub.uni-muenchen.de/27366/1/MPRA_paper_27366.pdf.

50. *See, e.g.*, *L.A. Tchr. Ratings*, L.A. TIMES, http://projects.latimes.com/value-added/ (last visited Jan. 19, 2012).

51. *See generally* Buddin, *supra* note 49. Derek Briggs and Ben Domingue of the University of Colorado conducted re-analysis of the L.A. Times data, showing that with modest model improvements (over the original Buddin model), some of the bias could be removed. *See* DEREK BRIGGS & BEN DOMINGUE, NAT'L EDUC. POL'Y CENTER, DUE DILIGENCE AND THE EVALUATION OF TEACHERS: A REVIEW OF THE VALUE-ADDED ANALYSIS UNDERLYING THE EFFECTIVENESS RANKINGS OF LOS ANGELES UNIFIED SCHOOL DISTRICT TEACHERS BY THE *LOS ANGELES TIMES* (Feb. 2011), *available at* http://nepc.colorado.edu/files/NEPC-RB-LAT-VAM_0.pdf.

between Asian race and math-teaching effectiveness being as strong as the negative association for black teachers.[52]

Some of these associations above are explained by related research by Stanford University Senior Fellow Eric Hanushek and Amherst College Professor Steven Rivkin, which shows measurable effects of the racial composition of peer groups on individual students' outcomes and explains the difficulty in distilling these peer effects from teacher effects.[53] Note that it is also likely that the above findings associated with teacher race are entangled with student race, so that black teachers are more likely to be in classrooms with larger shares of black students.[54]

All value-added comparisons are relative. They can be used for comparing one teacher to another in a school, teachers in one school to teachers in another school, or teachers in one district to those within other districts. The reference group becomes critically important when determining the potential for disparate impact of negative teacher ratings resulting from model bias. For example, employing a district-wide performance-based dismissal (or retention) policy in Los Angeles using Buddin's model would likely result in disproportionate layoffs of teachers in poor schools and black teachers of black students, while disproportionately retaining Asian teachers.[55] However, if one adopted the layoff policy relative to within-school rather than district-wide norms, because children are largely segregated racially and economically by neighborhoods and schools, the disparate effect might be lessened. The policy may neither be fairer nor better in terms of educational improvement, but racially disparate dismissals might be reduced.

Finally, because teacher value-added ratings cannot be disentangled entirely from patterns of student assignment across teachers within schools, principals may manipulate assignment of difficult and/or unmotivated students in order to compromise a teacher's value-added ratings, increasing the principal's ability to dismiss that teacher. This concern might

52. *Id.* 6, 7, 12, 16, and 14, respectively.

53. Hanushek & Rivkin, *supra* note 47.

54. Charles T. Clotfelter et al., *Who Teaches Whom? Race and the Distribution of Novice Teachers*, 24 ECON. EDUC. REV. 377 (2005).

55. *See* Buddin, *supra* note 49, at 12-16.

be mitigated by requirements for lottery-based student and teacher assignments. However, such requirements could create cumbersome student assignment processes that interfere with achieving the best teacher match for each child.

Whereas the problem of stability and error rates above are issues of "statistical error," the problem of non-random assignment is one of "model bias." Many value-added ratings of teacher effectiveness suffer from both large degrees of error and severe levels of model bias. The two are cumulative, not overlapping, problems. In fact, the extent of error in the measures may partially mask the full extent of bias.

F. Reduced Ability for Teachers to Receive VAM Ratings

In addition to the substantial concerns regarding "measurement error" and "model bias," which severely compromise the reliability and validity of value-added ratings of teachers as outlined above, in most public school districts, far fewer than half of certified teaching staff could even be assigned any type of value-added assessment score. While some reports suggest that as many as 30% might be assigned value added scores, when data demands are increased for applying more rigorous models, requiring more lagged student scores, thus reducing grade levels evaluated, these figures may drop significantly.[56] Existing standardized assessments typically focus on reading or language arts and math performance between grades three and eight.[57] Also, because baseline scores are required—ideally multiple prior scores to limit model bias—it becomes difficult to fairly rate third grade teachers.[58] By middle school or junior high, students are interacting with many more teachers, and it becomes more difficult to assign value-added scores to any one teacher.[59] When considering the

56. *See, e.g.*, CYNTHIA D. PRINCE ET AL., CENTER FOR EDUC. COMPENSATION REFORM, THE OTHER 69 PERCENT: FAIRLY REWARDING THE PERFORMANCE OF TEACHERS OF NONTESTED SUBJECTS AND GRADES (Aug. 2009), *available at* http://cecr.ed.gov/guides/other69Percent.pdf; CORCORAN ET AL., *supra* note 38. Briggs and Domingue's re-analysis of data from the L.A. Times study explains that even among the broad category of ratable teachers, only a relatively small share could actually be assigned ratings through more data rich models. *See* BRIGGS & DOMINGUE, *supra* note 51, at 22.

57. EVA L. BAKER ET AL., ECON. POL'Y INST, PROBLEMS WITH THE USE OF STUDENT TEST SCORES TO EVALUATE TEACHERS 16 (Aug. 29, 2010), *available at* http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf.

58. SCHOCHET & CHIANG, *supra* note 34, at 20.

59. *See* Koedel, *supra* note 44, at 682.

various support staff roles, specialist teachers, and teachers of elective and/or advanced secondary courses, value-added measures are generally applicable to only a small minority of teachers in any school district (less than 30%).[60] Thus, in order to make value-added measures a defined element of teacher evaluation in teacher contracts, one must have separately negotiated contracts for those teachers to whom these measures apply. Unfortunately, this is administratively cumbersome and potentially expensive for districts, especially in such difficult economic times.

Washington DC's IMPACT teacher evaluation system is one example that differentiates classes of teachers based on evaluation by including or excluding value-added measures.[61] While contractually feasible, this approach creates separate classes of teachers in schools and may have unintended consequences for educational practices, including increased tensions between non-value-added-rated teachers wishing to pull students of value-added-rated teachers out of class for special projects or activities.

IV. POSSIBLE LEGAL CHALLENGES BY TERMINATED TENURED TEACHERS

The previous section identified a number of technical issues that limit the effectiveness of VAM teacher evaluation plans. These technical problems make VAM teacher evaluation plans vulnerable to legal challenges by terminated tenured teachers.[62] Teachers may bring challenges pursuant to the Due

60. BAKER ET AL., *supra* note 57, at 12.

61. *See generally*, IMPACT GUIDEBOOKS, D.C. PUB. SCHS. (2011), *available at* http://dcps.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/IMPACT+Guidebooks.

62. Teachers who have been transferred from a non-VAM grade to a VAM grade might consider challenging the transfer. It is unlikely for these legal challenges to be successful, however. If the teacher is transferred to a grade in which she is certified, then she cannot make a challenge on Due Process Clause grounds. As the Pennsylvania Supreme Court explained in Smith v. Sch. Dist. of Darby, 130 A.2d 661, 665 (Pa. 1957): "A professional employee, under the tenure provisions of the Code, does not acquire a vested right to teach in any certain class or in any certain school. The only limitation on a school board's general power is that the work to which a professional employee is assigned be of a rank or class equivalent to that by which his permanent status was acquired and one for which he is qualified." (internal citations omitted). A teacher transferred from a non-VAM grade to a VAM grade might claim that the transfer constitutes a demotion in violation of a teacher tenure statute. *See, e.g.*, 24 PA. CONS. STAT. § 11-1151 (1963) ("there shall be no demotion of any

Process Clause, Equal Protection Clause, and Title VII. The remainder of this section elaborates on these options.

A. The Due Process Clause

The Due Process Clause of the Fourteenth Amendment provides that no state shall "deprive any person of life, liberty or property, without due process of law."[63] The Due Process Clause applies if the court determines that the plaintiffs have been deprived of a life, liberty, or property interest, whether procedurally or substantively. Procedural due process "(often summarized as 'notice and an opportunity to be heard'), is a right to a fair procedure or set of procedures before one can be deprived of property by the state."[64] Substantive due process "imposes limits on what a state may do regardless of what procedural protection is provided."[65]

1. Protected interest

To bring a Due Process Clause challenge, plaintiffs must first show that the government has infringed upon a life, liberty, or property interest. Plaintiffs may contend that a teacher evaluation system that relies significantly on student achievement data deprives them of a liberty interest. Deprivation of a liberty interest imposes a "stigma or other disability" that damages a person's standing in the community or forecloses a person's "freedom to take advantage of other employment opportunities."[66] On the other hand, "[a] statement that is basically one alleging conduct that fails to meet professional standards . . . does not impinge upon a liberty interest."[67]

---

professional employe [sic] either in salary or in type of position . . . without the consent of the employe [sic], or, if such consent is not received, then such demotion shall be subject to the right to a hearing before the board of school directors and an appeal"). However, a court is unlikely to view such a transfer as a "demotion" because there is no loss in salary or status. *See* Appeal of Santee, 156 A.2d 830 (Pa. 1959) (finding that a transfer from a secondary grade to an elementary grade was not a demotion because there was no difference in salary or prestige).

63. U.S. CONST. amend. XIV, § 1.

64. Seal v. Morgan, 229 F.3d 567, 574 (6th Cir. 2000).

65. Fournier v. Reardon, 160 F.3d 754, 757 (1st Cir. 1998). For more on the Due Process Clause, *see* Corinna B. Lain, *The Unexceptionalism of "Evolving Standards,"* 57 UCLA L. REV. 365 (2009).

66. Bd. of Regents of State Colls. v. Roth, 408 U.S. 564, 573 (1972).

67. Raposa v. Mead Sch. Dist. 46-1, 790 F.2d 1349, 1354 (8th Cir. 1986) (holding that school district's finding that teacher did not, inter alia, cooperate with other

It is doubtful that teachers terminated on the basis of their students' performance on achievement tests can establish a liberty interest. *St. Louis Teachers Union, Local 420 v. Board of Education of St. Louis*[68] supports this assertion. In this case, the St. Louis, Missouri school district adopted an evaluation system that called for certain teachers to be evaluated based on the California Achievement Test (CAT).[69] Plaintiffs who had received a preliminary score of "unsatisfactory" asserted that this rating violated their liberty interest.[70] The district court found that the school district's finding of "unsatisfactory" did not implicate a liberty interest because the district merely declared that the plaintiffs "did not meet professional standards for that year."[71] Such a finding did not damage the ability of the teachers to find employment elsewhere.[72]

Plaintiffs terminated on the basis of student performance might be able to establish a property interest, however. Property interests "are created and their dimensions are defined by existing rules or understandings that stem from an independent source such as state law."[73] In *Cleveland Board of Education v. Loudermill*, the Supreme Court found that tenured teachers had a property right to continued employment derived from the state tenure statute.[74]

2. Procedural due process

If plaintiffs can establish a protectable due process right, they might argue that a teacher evaluation policy that relies 50% or more on standardized tests, for example, violates procedural due process. Procedural due process is a flexible concept in that the procedural protections afforded to individuals depend on the demands of the particular situation. In *Mathews v. Eldridge*, the Supreme Court found that the level of procedural due process depends on three factors:

> First, the private interest that will be affected by the official action; second, the risk of an erroneous deprivation of such

---

teachers or failed to teach required subjects did not impinge upon a liberty interest).

68. 652 F. Supp. 425 (E.D. Mo. 1987).

69. *Id.* at 427.

70. *Id.* at 432.

71. *Id.*

72. *Id.*

73. Bd. of Regents of State Colls. v. Roth, 408 U.S. 564, 577 (1972).

74. 470 U.S. 532, 538-39 (1985).

> interest through the procedures used, and the probable value, if any, of additional or substitute procedural safeguards; and finally, the Government's interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirement would entail.[75]

With respect to the first factor, courts have consistently held that teachers have a considerable private interest in retaining employment.[76] With regards to the second factor, the risk of erroneous deprivation under value-added assessments is quite substantial; there is a 25% chance if a school district uses three years of data and a 35% chance if using one year of data that a teacher who is average would be identified as significantly worse than average and potentially fired. To base 50% or more of a teacher evaluation on such a flawed mechanism is therefore quite troubling.

The plaintiffs could argue that the governmental interest in hiring effective teachers does not outweigh these interests. It is unclear how the consideration of other quality measures such as "interim assessments results or evidence of student work"[77] will sufficiently mitigate the problem of overreliance on value-added assessments. Moreover, as to fiscal concerns, "[i]t is preferable to keep a qualified employee on than to train a new one."[78] Furthermore, "the employer shares the employee's interest in avoiding disruption and erroneous decisions,"[79] and it is quite possible that a termination process with such a high error rate might have a negative impact on the recruitment of qualified teachers.

3. Substantive due process

Plaintiffs might also claim that a teacher evaluation system that is based 50% or more on student achievement violates substantive due process. Substantive due process imposes limits on what a state may do regardless of what procedural

---

75. 424 U.S. 319, 335 (1976).

76. *Loudermill*, 470 U.S. at 543; Wash. Teachers' Union Local No. 6 v. Bd. of Educ. of D.C., 109 F.3d 774, 780 (D.C. Cir. 1997); Tex. Faculty Ass'n v. Univ. of Tex. at Dall., 946 F.2d 379, 384 (5th Cir. 1991).

77. COLO. REV. STAT. § 22-9-105.5(3)(a) (2010).

78. *Loudermill*, 470 U.S. at 544 (finding that terminated teachers are entitled to a pre-termination hearing).

79. *Id.*

protection is provided. A substantive due process analysis then requires that courts determine if the "life, liberty or property" interest in question is a fundamental right[80]—a right explicit or implicit in the federal constitution.[81] If a fundamental right is involved, the court reviews the legislative act using the strict scrutiny standard of review.[82] Under this standard of review, the burden is on the government to show that the legislative act is narrowly tailored to achieve a compelling governmental interest.[83]

If a fundamental right is not involved, the court reviews the legislative act under the more lenient rational basis standard of review. Under this standard, a violation of substantive due process occurs only if the legislative act is not rationally related to a legitimate state interest.[84] When an executive action or a specific act of a government official is challenged, substantive due process analysis requires courts to determine only if the executive action "shocks the conscience."[85]

If a tenured teacher is terminated due to the performance of her students on tests, then a rational basis analysis would be used because the termination occurred pursuant to a legislative act and a fundamental right is not implicated.[86] It is critical to point out that while plaintiffs have the burden of proof under rational basis review, the state could lose its case if its actions are arbitrary or irrational.[87] Plaintiffs in states where teacher

---

80. Dunn v. Fairfield Cmty. High Sch. Dist. No. 225, 158 F.3d 962, 965 (7th Cir. 1998).

81. San Antonio Indep. Sch. Dist. v. Rodriguez, 411 U.S. 1, 17 (1973).

82. *Id.*

83. Roe v. Wade, 410 U.S. 113, 155 (1973), *overruled in part on other grounds by* Gonzales v. Carhart, 550 U.S. 124 (2007).

84. FCC v. Beach Commc'n, Inc., 508 U.S. 307, 314 n.6 (1993).

85. Cnty. of Sacramento v. Lewis, 523 U.S. 833, 846-47 (1998).

86. No reported cases have addressed the question of whether tenured teachers can be terminated based on student achievement scores pursuant to substantive due process. Scheelhaase v. Woodbury Cent. Cmty. Sch. Dist., 488 F.2d 237 (8th Cir. 1973) (overturning lower court's finding that termination of teachers was arbitrary because the teacher was untenured at the end of her contract); St. Louis Teachers Union, Local 420 v. Bd. of Educ. of St. Louis, 652 F. Supp. 425 (E.D. Mo. 1987) (finding that tenured teachers stated a cause of action challenging salary decisions of certain teachers on the basis of student achievement test scores). It is critical to note that the Supreme Court has ruled that education is not a fundamental right. *See Rodriguez*, 411 U.S. at 1.

87. *See, e.g.*, Slochower v. Bd. of Higher Educ. of N.Y.C., 350 U.S. 551, 556 (1956) (quoting Wieman v. Updegraff, 344 U.S. 183, 192 (1952) ("constitutional protection does extend to the public servant whose exclusion [from public employment] pursuant to a statute is patently arbitrary or discriminatory"); *Roth*, 408 U.S. at 584 (quoting *Slochower*, 350 U.S. at 559) ("[T]he 'protection of the individual against arbitrary

evaluations are based 50% or more on value-added assessments might be able to make a tenable case for irrationality. As discussed, there is a 10% to 20% chance that a "good" teacher will be falsely identified as a "bad" teacher.[88] There are significant error rates, too: 25% with three years of data and 35% with one year of data. A court might find that the error rates are so extreme as to not be rationally related to any legitimate interest.

The high-stakes student testing case *Debra P. v. Turlington*[89] provides further support for a possible finding of irrationality. In *Debra P.*, Florida students challenged the constitutional validity of a state requirement that as a condition precedent to obtaining high school diplomas, students must pass a state test.[90] The students contended that such use of the test violated their due process rights. The United States District Court for the Middle District of Florida agreed, holding that the use of the test violated the students' due process property rights to a diploma due to a lack of adequate notice.[91]

The United States Court of Appeals for the Fifth Circuit remanded the case for further findings as to whether the high-stakes test "was a fair test of that which is taught in [Florida's] classrooms."[92] The circuit court stated that if the test covered content not actually taught in the state's classrooms, the test would violate substantive due process,[93] opining that "the state is obligated [under substantive due process] to avoid action which is arbitrary and capricious, does not achieve or even frustrates a legitimate state interest, or is fundamentally unfair."[94] The court went on to conclude that the high-stakes test may have violated substantive due process "in that it *may* have covered matters not taught in the schools of the state."[95] The record was "simply insufficient in proof that the test administered measures what was actually taught in the schools

---

action' . . . [is] the very essence of due process"); *id.* at 577 (protected property rights cannot be "arbitrarily undermined.").

88. SCHOCHET & CHIANG, *supra* note 34, at 12. *See supra* Part III.B.

89. 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd*, 644 F.2d 397 (5th Cir. 1981).

90. *Id.*

91. Debra P. v. Turlington, 644 F.2d 397, 402 (5th Cir. 1981).

92. *Id.* at 408.

93. *Id.* at 404.

94. *Id.*

95. *Id.* at 405.

of Florida."[96]

On remand, the district court found that the high-stakes test was instructionally valid and was therefore permissible under substantive due process. The circuit court affirmed the decision.[97] It rejected the notion that the state had to demonstrate that the test covered materials actually taught in the classroom because there were no accepted standards for deciding whether a test was instructionally valid,[98] concluding instead that there was adequate evidence of its being instructionally valid.[99] Among other things, the court was impressed by the state's efforts to provide remedial instruction to students who needed extra help mastering the skills on the test,[100] as well as by a student survey finding that 90 to 95% of the students believed that they had been taught the test skills.[101]

If a court were to adopt the *Debra P.* approach, it might determine that a teacher evaluation program based 50% or more on a value-added assessment, having a significant 25% error rate with three years of data and 35% error rate with one year of data, violates substantive due process. The crucial question would be whether it is fundamentally unfair to base the decision to terminate a teacher on an assessment with such high error rates. The circuit court in *Debra P.* found fundamental fairness only because the state of Florida provided remediation for those students who failed the examination.[102]

Yet there is an important distinction between the high-stakes test analyzed in *Debra P.* and the teacher evaluation programs analyzed in this Article. In *Debra P.*, the Fifth Circuit found that there were "no accepted educational standards for determining" what constitutes instructional validity[103] and so the state's task in establishing the validity of the high-stakes test was made relatively easy. By contrast, the Economic Policy Institute states that "there is broad agreement

96. *Id.*
97. Debra P. v. Turlington, 730 F.2d 1405, 1406 (11th Cir. 1984).
98. *Id.* at 1409.
99. *Id.* at 1411.
100. *Id.*
101. *Id.*
102. *Id.* at 1411, 1416.
103. *Id.* at 1412 n.4.

among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions, even when the most sophisticated statistical applications such as value-added modeling are employed,"[104] making it more difficult for VAM to receive court approval.

Because of this consensus, a court may find that sole reliance on a value-added assessment model would be fundamentally unfair. Thus, the question the court may decide to address under *Debra P.* is whether the other measurements and remedial policies sufficiently counter the value-added-assessment system's inherent irrationality. All of the state policies analyzed in this Article rely on remediation of teachers who are deemed ineffective. However, as explained, it may be more difficult to provide remediation for teachers whose students fail to meet achievement benchmarks. A teacher could do a wonderful job but not see academic gains because of the socioeconomic, racial, and ability composition of her class. In other words, remediation may be less effective because of factors completely out of control of the teacher. Additionally, a teacher who is indeed effective, yet forced to undergo remediation due to a false identification based on the VAM, might face consequent stigma.

B. The Equal Protection Clause

The Equal Protection Clause provides in pertinent part: "No State shall . . . deny to any person within its jurisdiction the equal protection of the laws."[105] It "does not forbid classifications. It simply keeps governmental decisionmakers from treating differently persons who are in all relevant respects alike."[106]

Courts generally use three levels of analysis for Equal Protection Clause challenges. Governmental classifications that implicate a fundamental right or target a suspect class are subject to strict scrutiny, which requires a narrow tailoring to satisfy a compelling governmental interest. Quasi-suspect classifications such as gender and illegitimacy are subject to

104. BAKER ET AL., *supra* note 57, at 2.

105. U.S. CONST. amend. XIV, § 1.

106. Nordlinger v. Hahn, 505 U.S. 1, 10 (1992).

intermediate scrutiny, requiring substantial relation to an important governmental interest. All other classifications are subject to a rational basis analysis, meaning that they will be constitutional as long as they are rationally related to a legitimate governmental interest.[107]

Courts reviewing equal protection claims of tenured teachers terminated because of their students' academic performance will probably use a rational basis analysis because: (1) no fundamental right is implicated; and (2) terminated tenured teachers are not a suspect or quasi-suspect class. In *Debra P.*, the circuit court held that a high-stakes student test that fails to cover the material within the curriculum could not be rationally related to a legitimate governmental interest. In other words, "[i]f the test is not fair, it cannot be said to be rationally related to a state interest."[108] It might be possible that the error rates of value-added estimates might make teacher evaluation policies that are 50% or more reliant on such tests "too unfair" or arbitrary to satisfy the rational basis test, similar to the discussion under the Due Process Clause.

C. Title VII

The third section of this Article presented research indicating that: (1) black students tend to fare worse on standardized tests than white students; and (2) black teachers are more likely to work in schools of low-income black students.[109] Thus, it follows that black teachers are more likely to be dismissed on the basis of poor value-added test scores. This is especially true if states adopt teacher evaluation systems that rely 50% or more on student standardized test scores.

The potential racial impact of such systems may make them vulnerable to challenges under Title VII of the Civil Rights Act of 1991, which makes it unlawful for an employer "to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race."[110] There are two types of

107. *Id.*

108. Debra P. v. Turlington, 644 F.2d 397, 406 (5th Cir. 1981).

109. *See supra* Part III.E.

110. 42 U.S.C. § 2000e-2(a)(1) (2010).

Title VII challenges: (1) disparate treatment, which deals with purposeful discrimination; and (2) disparate impact, which addresses employment policies that are "fair in form but discriminatory in operation."[111] While no doubt disparate treatment cases may occur, it is much more likely that teacher evaluation policies will be enacted without any provable racial animus. For this reason, this Article analyzes racial effects under a disparate impact analysis.[112]

Courts apply a three-part, burden-shifting analysis for Title VII disparate impact claims.[113] First, the plaintiffs must establish a prima facie case, showing that a challenged practice has an adverse impact on a minority group.[114] Once the plaintiffs have established a prima facie case, the burden shifts to the employer to show that the employment practice in question has a "manifest relationship to the employment";[115] in other words, the employer has to show a "business justification."[116] If the employer satisfies this requirement, the burden then shifts to the plaintiffs to establish that less discriminatory alternatives exist.[117]

---

111. Connecticut v. Teal, 457 U.S. 440, 455-56 (1982) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971)).

112. A purposeful discrimination challenge could also be made pursuant to the Equal Protection Clause. If plaintiffs could establish that purposeful discrimination has occurred, the case would be analyzed under strict scrutiny, which is the most difficult level of analysis for the government to overcome. If the plaintiffs can prove racially discriminatory purpose for value-added designs or ratings, either through direct or circumstantial evidence, courts will not defer to the legislature. Such evidence could include "a clear pattern, unexplainable on grounds other than race, [that] emerges from the effect of the state action even when the governing legislation appears neutral on its face." Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 266 (1977). Other evidence of discriminatory purpose could come from the "historical background" of the enactment, design and/or use of the VAM, especially "if it reveals a series of official actions taken for invidious purposes." *Id.* at 267. However, it would be extremely difficult to find such evidence. *See generally Arlington Heights*, 429 U.S.; Washington v. Davis, 426 U.S. 229 (1976) (highlighting the difficult nature of Equal Protection cases founded on racially discriminatory impact). Due to the difficulty in proving racial animus and the fact that in current times government officials are less likely to leave a paper or electronic trail of such animus when crafting policy or legislation, it would be better for plaintiffs to focus on disparate impact Title VII challenges since proof of discriminatory intent/purpose is not required.

113. *See, e.g.*, 42 U.S.C. 2000e-2(k) and Gulino v. N.Y. State Educ. Dep't, 460 F.3d 361, 382 (2d Cir. 2006).

114. *Gulino*, 460 F.3d at 382.

115. *Id.*

116. *Id.*

117. *Id.* In Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659 (1989), the Supreme Court held that in Title VII disparate impact, "'the ultimate burden of proving that discrimination against a protected group has been caused by a specific

Plaintiffs should have no difficulty establishing a prima facie case of disparate impact. As noted above, black teachers receive lower value-added scores than white teachers in at least some existing publicly disclosed teacher effectiveness (VAM) reports.[118] Additionally, even though relatively small, there is a negative correlation between the number of black students a teacher has and the teacher's VAM scores,[119] and black teachers are more likely to work in majority-black schools.[120] Furthermore, with research showing that it is difficult to separate the effects of peer racial composition from teacher effects on student performance, the racially disparate impact of VAM cannot be trivialized.[121]

Assuming that black teachers are able to establish a prima facie case, the burden would then shift to the defendants to establish a business justification for the teachers' dismissal. The defendants would centrally argue as a business justification the need to ensure that ineffective teachers are not teaching students, for teacher quality determines student performance. However, the fact that value-added tests are riddled with error rate problems may make it difficult for the defendant state to establish such a manifest relationship.

If the school district gets beyond the "business justification" hurdle, plaintiffs might suggest that a less racially discriminatory alternative would be to explicitly include indicators of the racial mix of students in the class as part of the teacher evaluation model. Doing so would hypothetically compare teachers of classrooms of children where racial composition of classrooms is statistically (albeit not practically) equalized. That is, teachers serving classrooms of predominantly black students would be compared against teachers serving classrooms of the same. In practice, many value-added models like the *Los Angeles Times* model avoid

---

employment practice remains with the plaintiff at all times.'" (quoting Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 997 (1989) (emphasis in original)). In 1991, Congress responded to *Wards Cove* by codifying the disparate impact analysis prior to *Wards Cove*. 42 U.S.C. § 2000e-2(k)(1)(C) (2011). However, Congress failed to make similar changes to other employment discrimination statutes, such as the Age Discrimination in Employment Act (ADEA). 29 U.S.C. § 623 (2011). In Smith v. City of Jackson, 544 U.S. 228, 240 (2005), the Court held that pre-*Wards Cove* analysis applies to ADEA claims.

118. *See* Buddin, *supra* note 49.

119. *Id.*

120. Clotfelter et al., *supra* note 54, at 377.

121. Hanushek & Rivkin, *supra* note 47.

including race variables despite knowing their importance in resolving key modeling bias problems. While one can be reasonably sure that typical value-added models do not entirely mitigate racially disparate effects, even the most thorough value-added models cannot guarantee such results.

V. RANDOMNESS: THE BITTERSWEET VAM PILL

The previous section noted that VAM for teacher evaluations that rely 50% or more on teacher scores may be subject to a variety of legal challenges. These challenges arise from the fact that VAM fails to adjust sufficiently for the differences in student populations. In order for VAM to work, it would be necessary to tackle the non-randomness problem in student assignments and teaching conditions.[122] For example, teachers might consider arguing for a "randomized student assignment clause" in their collective bargaining agreements ("CBA") to require all students in any given grade level to be randomly assigned to classrooms and schools, stratified by student population characteristics including disabilities (by type), language proficiency, socio-economic status, and race.

Teachers might also argue for more detailed "comparable conditions" clauses in their CBAs. These comparable conditions clauses would specify precisely the number of children to be taught per class or section and the numbers and types of children by various classifications (as per the random assignment system). Additionally, teachers could argue for comparable facilities, accommodations and other resources, including air quality, heating, cooling, lighting quality, materials, supplies, and equipment, and any other factors that may bias teacher "effectiveness" ratings. Furthermore, where teachers work in teams with students, requirements of rotating schedules to ensure a distribution of "time of day" for students in classes or sections might be considered.

However, while randomized assignments would help address some of the concerns discussed in this article, the potential for unintended consequences exists. For instance, completely random assignment of students and random matching of students to teachers and classrooms removes the option for principals to work with teachers to determine the

122. BAKER ET AL., *supra* note 57, at 9-11.

best match for each child and eliminates the possibility for teachers wishing to dedicate themselves to assisting more difficult children. Establishing comparison groups so that highly successful and predominantly white schools or districts must dismiss x percent of their teachers yearly simply to erase the racially disparate effect of dismissing x percent of teachers in poor minority districts based on their school- or district-level norms is equally absurd.

Moreover, these proposed contractual solutions might only apply to a relatively small share of teachers in the system, given the fact that only about 20% of teachers can be linked directly to student performance measures in reading and math.[123] Additionally, in most studies of the stability of value-added measures, math performance measures have been found much more stable than reading performance measures, and reading performance measures are much more strongly influenced by student learning outside the control of teachers over the summer.[124] As such, contracts for teachers evaluated via value-added measures must generally differ from those of other teachers, and it may be necessary to include different protections for teachers of reading than math to account for different levels of model error and different effects of student sorting based on differential summer learning patterns. For example, while fall-spring assessments are more appropriate than annual assessments for determining teacher effects in either reading or math, it would appear more important to include contractual requirements for fall-spring assessments for rating teachers of reading.

The lack of randomness in student assignments that typifies VAM effectively ensures that the system will remain beset with other problems. The different pressures placed on teachers of reading and math between grades three and eight and other teacher specialists working with the same students may create unintended curricular consequences. For instance, a core-content-area teacher might refuse release time for those students that would most help that teacher improve her value-added ratings while encouraging release time or classroom removal for more disruptive students that might negatively

123. *See supra* Part III.F, pointing out that standardized assessments currently used mostly focus on language arts or reading and math performance between grades three and eight.

124. BILL & MELINDA GATES FOUND., *supra* note 29.

impact her ratings. Furthermore, re-assignment of teachers into and out of value-added-rated contractual categories, including moving a teacher from second grade (not rated) to third or fourth, may become much more cumbersome and even lead to illicit backroom deals between teachers and administrators as teachers seek to avoid falling into value-added-rated categories. This could severely compromise the integrity of the educational process and, indeed, student achievement. Perhaps the least reasonable solution is simply to test everything from kindergarten art work to the high school jazz ensemble's performance in order to apply value-added ratings to teachers responsible for each aspect of student work as they pass through the school system. While seemingly absurd, states including Tennessee have established committees to explore this and other equally problematic possibilities, including the evaluation of teachers in music and art according to school average achievement growth in tested subject areas or of counselors by student discipline referral rates.[125] Alternatively, schools might simply choose to discontinue offering anything that is not presently tested.

VI. CONCLUSION

In response to complaints about teacher evaluation programs, several states have adopted VAM policies. Three states have gone so far as to base 50% of their teacher evaluation policies on VAM. This Article advises against such reliance on VAM because of the attendant technical and legal problems. Random assignments provide the best possible avenue for alleviating some of the problems identified, but even then only as a bittersweet pill given the potential unintended consequences. Unless randomness can be incorporated into VAM models, the basic notion of fairness demands that states refrain from relying on a flawed model with such high error rates in determining the fate of their teachers. A 25% chance of error is not acceptable, while a 35% chance of erroneous deprivation is unconscionable. Since VAM is in its incipiency, states still have an opportunity to incorporate the above suggestions to make the system fair and less susceptible to

125. *Tennessee Teacher Evaluation Advisory Committee*, TN.GOV, http://www.tn.gov/education/TEAC.shtml (last visited Jan. 20, 2012) (see meeting minutes for details of options considered).

challenges, for, as noted author Orlando A. Battista once stated, "[a]n error doesn't become a mistake until you refuse to correct it."[126]

126. *Quotes by Orlando A. Battista*, QUOTEWORLD.ORG, http://www.quoteworld.org/authors/orlando_a__battista (last visited Jan. 20, 2012).