The Report committee for Leila Melanie Melhem

Certifies that this is the approved version of the following report:

Are Value-Added Models for High-Stakes Teacher Accountability Arbitrary and Capricious?

Supervisor: _____
Cynthia Osborne

Co-Supervisor:_____
Norma V. Cantú

Are Value-Added Models for High-Stakes Teacher Accountability Arbitrary and Capricious?

by

Leila Melanie Melhem, B.A., M.T.

Report

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degrees of

Master of Public Affairs

and

Doctor of Jurisprudence

The University of Texas at Austin

May 2012

Abstract

Are Value-Added Models for High-Stakes Teacher Accountability Arbitrary and Capricious?

by

Leila Melanie Melhem, J.D.;M.P.Aff.

The University of Texas at Austin, 2012

SUPERVISORS: Cynthia Osborne, Norma V. Cantú

Value-added models are complex statistical formulas that aim to isolate the effect a teacher has on student learning. States and districts across the nation are adopting laws and policies that will evaluate teachers, in part, using the results provided by value-added models. In many states and districts, these evaluations will be used to inform high-stakes decisions about teacher salary and retention. However, value-added models are imperfect tools for assessing teacher effectiveness, and many scholars have argued that they are not appropriate for use in high-stakes decisions.

This Article provides a brief history of the use of value-added models in public education and summarizes the major criticisms of using value-added models. In this

context, the Article analyzes and evaluates the extent to which substantive due process claims brought by teachers adversely affected by the results of value-added models will be successful. The Article concludes that while the system as a whole is rationally related to the objective of improving the overall effectiveness of the teaching workforce, in certain cases, individual teachers will be able to successfully claim that the results of their value-added model led to a termination that was arbitrary and capricious. Finally, the paper offers some recommendations to states and school districts on how to implement an evaluation system using value-added models to avoid substantive due process violations.

Table of Contents

PART I. INTRODUCTION

Value-added modeling is becoming an increasingly popular way to evaluate public school systems and public school teachers across the country. More and more, school districts are using these models to hold individual teachers accountable by using the results as a factor in determining who should receive bonuses and who should be fired.[1] Value-added models are complex statistical formulas that aim to isolate the effect a teacher has on student learning. There are many versions of value-added models; variations depend on the number and type of variables used in the model, and the weight assigned to each variable. As states and districts across the country each adopt their own model, there will also be variations in how the models are implemented. Because, at this stage, value-added models are imperfect tools for assessing teacher effects on student learning, states and school districts should be wary of the potential legal implications that may arise from their use.

This Article seeks to provide the reader with a brief background of value-added modeling, identify and evaluate the potential for substantive due process challenges that could arise through the use of value-added modeling in making high-stakes decisions about teachers, and provide recommendations that states and school districts can use to implement value-added models to avoid legal challenges, or at least, avoid adverse judicial decisions resulting from those challenges. Dealing with the various, specific permutations that value-added models and the implementation of such models is beyond

[1] Sam Dillon, *Formula to Grade Teachers' Skill Gains Acceptance, and Critics*, N.Y. TIMES, Aug. 31, 2010 at A1.

the scope of this Article. Instead, the Article will deal only with the generalized problems presented by the use of value-added models.

Part II of this Article provides a brief history of the use of value-added modeling in public education. Part III summarizes the most significant criticisms of value-added models and their implementation. Part IV of this Article describes the basic substantive due process doctrine and lays out the framework for determining whether teachers will have a substantive due process claim arising from their termination. The Article then analyzes whether using value-added models to make high-stakes decisions about teachers is arbitrary and capricious and under what circumstances. Part V of the Article offers recommendations on how school districts can best implement value-added models to avoid legal challenges and briefly concludes.

PART II. VALUE-ADDED MODELING'S BACKGROUND

A. WHY THE INTEREST IN APPLYING VALUE-ADDED MODELS TO TEACHERS?

Research consistently shows that teachers have a large and lasting impact on student achievement.[2] In fact, many researchers believe that among all the resources found in schools, teachers have the largest effect on student achievement.[3] It makes good sense then to want to differentiate which teachers improve student achievement most and which teachers add the least to student growth.

Generally, however, the current system of evaluating teachers in most schools and school districts does a poor job of differentiating among teacher quality.[4] In *The Widget Effect*, the New Teacher Project found that teacher effectiveness "is not measured, recorded, or used to inform decision-making in any meaningful way" in most schools across the country. In districts that used binary evaluation ratings (e.g., satisfactory/unsatisfactory), more than ninety-nine percent of teachers were rated satisfactory.[5] Districts that had more than two categories still had less than one percent of teachers rated in the lowest category.[6]

---

[2] *E.g.,* Raj Chetty, John N. Friedman, & Jonah E. Rockoff, *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood* (Nat'l Bureau of Econ. Research, Working Paper No. 17699, 2011); B. Rowan, R. Correnti, & R.J. Miller, *What Large-Scale Survey Research Tells Us About Teacher effects on Student Achievement: Insights from the Prospects Study of Elementary Schools*, 104 Teachers College Record 1525 (2002).

[3] D. Boyd, et al., *How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement*, 1 Educ. Finance and Policy 176, 177 (2006).

[4] THE NEW TEACHER PROJECT, THE WIDGET EFFECT: OUR NATIONAL FAILURE TO ACKNOWLEDGE AND ACT ON DIFFERENCES IN TEACHER EFFECTIVENESS (2009) *available at* http://widgeteffect.org/.

[5] *Id.* at 6.

[6] *Id. at* 3.

Under the 2001 No Child Left Behind Act, "highly qualified" teachers were defined as those who had a bachelor's degree, a state license, and a college major or a master's degree in the subject matter taught.[7] Research has shown, however, that these qualifications are not "strongly predictive of student outcomes on standardized tests."[8] Rather, research has shown that the best predictor of a teacher's effectiveness is his or her past success in the classroom.[9]

Value-added models offer the promise of isolating teacher contributions to student growth in academic achievement. Although value-added models vary, they generally work by comparing a student's actual performance on a standardized test to the student's predicted score, which is based on that student's past performance and other characteristics believed to influence student learning. Teachers who produce more growth than expected are said to "add value."[10] Generally, value-added models show that there is, in fact, "substantial variation in teacher quality as measured by the value added to achievement or future academic attainment or earnings."[11] However, beyond this showing, value-added models suffer from many technical flaws; isolating a teacher's unique contribution to student learning is extremely difficult and complex.[12] Consequently, researchers "mostly concur" that using value-added modeling to make

[7] No Child Left Behind Act of 2001, 20 U.S.C.A. § 9101(23).
[8] SEAN P. CORCORAN, 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 1 (Annenberg Institute for School Reform at Brown University, 2010).
[9] THE NEW TEACHER PROJECT, TEACHER EVALUATION 2.0, Preface (2010) *available at* http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf?files/Teacher-Evaluation-Oct10F.pdf.
[10] U.S. DEP'T OF EDUC., MEASURING TEACHER EFFECTIVENESS USING GROWTH MODELS: A PRIMER 1 (2011).
[11] ERIC A. HANUSHEK & STEVEN G. RIVKIN, GENERALIZATIONS ABOUT USING VALUE-ADDED MEASURES OF TEACHER QUALITY 1 (2010).
[12] CORCORAN *supra* note 8, at 4.

high-stakes decisions "should be pursued only with great caution."[13] Nevertheless, states and school districts across the country have been rapidly moving toward using the results from value-added models to make high-stakes decisions regarding teacher retention and pay.

B. THE RISING USE OF VALUE-ADDED MODELS IN PUBLIC SCHOOL SYSTEMS

Using value-added modeling in public school systems is a relatively recent phenomenon. Tennessee has used value-added modeling since 1992, [14] Ohio began mandating its use in 2003,[15] and Pennsylvania has required its use for all districts since 2005.[16] At first, these states used value-added modeling only to rate the performance of their schools and districts overall, and to help teachers make instructional decisions to ensure the academic growth and achievement of their students.[17] Value-added modeling was not used to evaluate individual teachers or used to make high-stakes decisions about individual teacher employment or pay.

The use of value-added modeling to evaluate individual teachers increased, at least in part, due to the American Recovery and Investment Act of 2009 (ARRA), which

---

[13] ECONOMIC POLICY INSTITUTE BRIEFING PAPER, PROBLEMS WITH THE USE OF STUDENT TEST SCORES TO EVALUATE TEACHERS 7 (2010) [hereinafter EPI BRIEFING PAPER].
[14] Theodore Hershberg, *Value-Added Assessment*, Operation Public Education, The Center for Greater Philadelphia, http://www.cgp.upenn.edu/ope_value.html.
[15] Theodore Hershberg, *Value-Added Assessment in Ohio*, Operation Public Education, The Center for Greater Philadelphia, http://www.cgp.upenn.edu/ope_ohio.html.
[16] Theodore Hershberg, *Value-Added Assessment in Pennsylvania*, Operation Public Education, The Center for Greater Philadelphia, http://www.cgp.upenn.edu/ope_pa.html.
[17] Pa. Dep't of Educ., *Pennsylvania Value Added Assessment System*, http://www.portal.state.pa.us/portal/server.pt/community/pa_value-added_assessment_system_(pvaas)/8751; Tn. Dep't of Educ., *Tennessee Value-Added Assessment System – TVAAS*, http://www.tn.gov/education/assessment/test_results.shtml; Oh. Dep't of Educ., *Guide to Understanding Ohio's Accountability System 2009-2010*, http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=117&ContentID=91231&Content=91251

President Obama signed into law on February 17, 2009.[18] The ARRA encouraged states to implement educational reform by awarding grants from a $4.35 billion Race to the Top Fund to the states that best "creat[ed] the conditions for education innovation and reform" in four core areas:

- Adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy;
- Building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction;
- Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and
- Turning around our lowest-achieving schools.[19]

State applications for Race to the Top funds were scored on selection criteria worth 500 points.[20] Fifty-eight points, or over ten percent of the total points awarded, were awarded for plans to improve teacher and principal effectiveness based on performance.[21] Another forty-seven points were distributed for implementing and using data systems to support instruction, including the use of longitudinal data systems.[22] Longitudinal data systems that link teachers to students enable policymakers to base teacher evaluations on student achievement gains.[23]

---

[18] American Recovery and Reinvestment Act of 2009, Pub.L. No. 111-5 (H.R. 1), 123 Stat. 115 (Feb. 17, 2009) as amended by Pub.L. No. 111-8 (H.R. 1105), the Omnibus Appropriations Act, 2009, Division A, Section 523, 123 Stat. 524 (Mar. 11, 2009).
[19] U.S. DEP'T OF EDUCATION, RACE TO THE TOP PROGRAM EXECUTIVE SUMMARY, 2 (November 2009).
[20] *Id.* at 3.
[21] *Id.*
[22] *Id.*
[23] JENNIFER L. STEELE, LAURA S. HAMILTON, & BRIAN M. STECHER, INCORPORATING STUDENT PERFORMANCE MEASURES INTO TEACHER EVALUATION SYSTEMS iii, (RAND Corporation, 2010).

Delaware and Tennessee, the only two winners of the first round of Race to the Top,[24] both incorporated value-added methods to evaluate teacher performance as part of their applications. All ten winners of the second round of Race to the Top included value-added methods to evaluate teacher performance, although the stakes attached to the evaluation and the percentage of the evaluation based on the value-added model varied.[25]

As of September 2011, twenty-three states required that teacher evaluations include "objective evidence of student learning in the form of student growth and/or value-added data."[26] Five states have adopted legislation or regulations that specifically require that student achievement and/or student growth "significantly" inform teacher evaluations, while eleven states and the District of Columbia Public Schools (DCPS) make student achievement/growth the "preponderant" criterion in teacher evaluations.[27] For example, Tennessee requires that fifty percent of a teacher's evaluation come from student achievement data, of which thirty-five percent must come from student growth data from the Tennessee Value-Added Assessment System (TVAAS).[28] Similarly, in DCPS, value-added

[24] Justin Hamilton, *Delaware and Tennessee Win First Race to the Top Grants*, ED.gov, U.S. Dep't of Education, Mar. 29, 2010.
[25] See http://www2.ed.gov/programs/racetothetop/index.html for individual state applications for Race to the Top Funds.
[26] National Council on Teacher Quality, State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies, ii (Oct. 2011) [hereinafter NCTQ].
[27] *Id.* at 6.
[28] *Id.* at 16.

data make up fifty percent of a teacher's evaluation score.[29] New York, which only requires a "significant," not "preponderant," portion of a teacher's total evaluation be based on student growth data, requires that twenty-five percent of a teacher's evaluation be based on student growth on state assessments.[30]

The consequences tied to teacher evaluations that are based on value-added student growth vary widely. In thirteen states, teachers are eligible for dismissal based on teacher evaluations that are tied to student performance, and ten of these are states that use student growth as the preponderant criterion.[31] Illinois, Indiana, and New York use student achievement as a significant factor in teacher evaluations and require teachers to be dismissed on the evaluation results.[32] Several other states require teachers to be dismissed on the basis of evaluations, but student achievement/growth is not a factor in their evaluations.[33] The chart below[34] (Figure 1) describes policies for (1) assisting teachers who receive poor evaluations and (2) dismissing ineffective teachers for states that base at least a significant portion of the teacher evaluation on student achievement/growth measures and require teachers to be eligible for dismissal based on their evaluations.

---

[29] *Id.*
[30] *Id.*
[31] *Id.* at 6. The ten states that require teachers to be eligible for dismissal based on evaluations in which student achievement/growth is the preponderant criterion are Colorado, Delaware, DCPS, Florida, Louisiana, Michigan, Nevada, Oklahoma, Rhode Island, and Tennessee. *Id.*
[32] NCTQ, *supra* note 26, at 6.
[33] *Id.*
[34] Excerpted from the National Council on Teacher Quality, State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies. A column showing the actual proportion of the evaluation based on student/growth achievement measures has been added.

Figure 1

STATE	VAM % of evaluation	Policy for assisting teachers who receive poor evaluations	Policy for dismissing ineffective teachers
CO	50	Each teacher must be provided with an opportunity to improve effectiveness through a teacher development plan. Teachers who object to a rating have an opportunity to appeal. For non-probationary teachers, the district must develop a remediation plan that includes professional development opportunities. The teacher must be given a "reasonable period" to improve.	Colorado specifically identifies classroom ineffectiveness as grounds for dismissal. For teachers who receive a performance rating of ineffective, the evaluator shall either make additional recommendations for improvement or may recommend dismissal. Non-probationary teachers who receive two consecutive ineffective ratings return to probationary status and have one year to improve or face termination.
DE	50	Teachers who receive an overall rating of needs improvement or ineffective on the summative evaluation, or a rating of unsatisfactory on any appraisal component regardless of the overall rating, must be put on an improvement plan.	A teacher cannot be rated "effective" overall if the student growth expectations for the teacher's students are not met. Teachers with two consecutive years of ineffective ratings or who earn a combination of ineffective and unsatisfactory ratings for three consecutive years are considered to have a pattern of ineffective teaching and are eligible for dismissal.
D.C.	50 (individual) 5 (school wide)	Those who are rated minimally effective are encouraged to take advantage of professional development opportunities provided by DCPS.	DCPS ensures that teacher ineffectiveness is grounds for dismissal. Individuals who receive ineffective ratings are "subject to separation from the school system."
FL	50, unless there are fewer than 3 years of data for a teacher, then 40	If a teacher receives an unsatisfactory evaluation, the evaluator must make recommendations as to specific areas of unsatisfactory performance and provide assistance in helping to correct deficiencies within a prescribed period of time.	Teacher ineffectiveness is grounds for dismissal. All new teachers are placed on annual contracts and the state requires that such contracts are not renewed if a teacher's performance is unsatisfactory. An annual contract may not be awarded if the teacher has received two consecutive annual performance evaluation ratings of unsatisfactory, OR two annual performance ratings of unsatisfactory within a three-year period, OR three consecutive annual performance evaluation ratings of needs improvement or a combination of needs improvement and unsatisfactory.
IN	"significant;" not specified	Not specified	Indiana ensures that teacher ineffectiveness is grounds for dismissal. A tenured teacher reverts to probationary status if the teacher has received a rating of ineffective on an evaluation and can be subject to contract cancellation for a rating of ineffective in the year immediately following the teacher's initial rating of ineffective.
LA	50	Any teacher not deemed effective will be placed in an intensive assistance program and then must be formally re-evaluated. Program must include an expected time line for achieving objectives; must not exceed two years.	If at the end of intensive assistance program, a teacher does not complete the program or is still deemed ineffective based on evaluation, the school district is allowed to initiate termination proceedings.
MI	25 (2013-14) 40 (2014-15) 50 (2015-16)	Teacher must be given "ample opportunities for improvement."	Classroom ineffectiveness is grounds for dismissal. If a teacher is rated as ineffective on 3 consecutive annual year-end evaluations, the district shall dismiss the teacher.
NY	25 (state assessments) 15 (local measure)	If a teacher is rated developing or ineffective, the school district is required to develop and implement a teacher or principal improvement plan.	Tenured teachers with a pattern of ineffective teaching or performance, defined as two consecutive annual ineffective ratings, may be charged with incompetence and considered for termination through an expedited hearing process.
OK	35	All teachers who receive ratings of needs improvement or ineffective must be placed on comprehensive remediation plans and provided with instructional coaching.	Oklahoma ensures that teacher ineffectiveness is grounds for dismissal. Teachers rated as ineffective for two consecutive years; needs improvement for three years; or for those who do not average at least an effective rating over a five-year period shall be dismissed.
TN	35	Not specified	Tennessee explicitly makes teacher ineffectiveness grounds for dismissal. Tennessee specifies that tenured teachers who receive two consecutive years of below expectations or significantly below expectations performance ratings are returned to probationary status, making them eligible for dismissal.

In addition, Florida has eliminated teacher tenure, while other states have made earning tenure contingent on receiving certain scores from the value-added model.[35] To earn tenure, probationary teachers must earn three consecutive "effective" ratings in Colorado, show "satisfactory growth" in two out of three years in Delaware, and wait a minimum of five years in Michigan, with a rating of at least "effective" for the three most recent annual performance evaluations.[36] Colorado, Florida, Michigan, and Indiana have eliminated or restructured "last in, first out" policies for reduction in force, putting a greater emphasis on teacher performance.[37] Louisiana and Rhode Island are tying renewal of teacher certifications to value-added performance indicators.[38] Florida is tying teacher compensation to teacher performance; highly effective teachers will receive a salary increase greater than they could through any other salary schedule adopted by the district, and effective teachers will receive a salary increase equal to 50-75% of what the highly effective teachers receive.[39]

Nevertheless, only a few states, including Rhode Island, Tennessee, and Delaware, have already implemented value-added based teacher evaluation systems. Only DCPS has applied consequences to teachers based on the results of their evaluations, although only "to those teachers whose

[35] *Id.* at 25.
[36] *Id.* at 25-26.
[37] *Id.* at 26.
[38] *Id.* at 27.
[39] NCTQ, *supra* note 26, at 27.

evaluations have shown either exceptional or very poor performance."[40] For example, Michelle A. Rhee, the former schools chancellor in Washington, D.C., fired about twenty-five teachers in the summer of 2010 "after they rated poorly in evaluations based in part on a value-added analysis of [test] scores."[41] In 2011, 206 public school teachers in D.C., approximately five percent of the public school teaching force, were fired on the basis of poor value-added scores.[42]

Although only a few states have actually implemented teacher evaluations based on value-added models, many individual school districts have taken it upon themselves to implement value-added teacher evaluations. Houston Independent School District (Houston ISD), for example, has been using value-added data since 2007 to determine performance bonuses for its teachers.[43] In January 2010, Houston ISD awarded approximately $40.4 million in performance bonuses based on results from the 2008-2009 school year.[44] Additionally, in 2010 the Houston ISD school board approved a plan to terminate teachers based on results from its value-added model. Dallas ISD has been using value-added measures since 1992, but only to identify high-performing schools. Beginning in 2007, Dallas ISD implemented the "Principal and Teacher Incentive Pay" program which awarded eligible

[40] *Id.* at 1.
[41] Dillon, *supra* note 1.
[42] Bill Turque, *More than 200 D.C. Teachers Fired*, WASH. POST, July 15, 2011.
[43] Ericka Mellon, *Statistical Tool for Rating Faculty Could Be Adopted Today By HISD*, HOUSTON CHRONICLE, May 11, 2011.
[44] HOUSTON INDEP. SCH. DIST., ANALYSIS FREQUENTLY ASKED QUESTIONS 1, *available at* http://www.houstonisd.org/HISDConnectEnglish/Images/PDF/ValueAdded_FAQ_0209.pdf.

teachers $1,600-$4,000 incentives based upon individual value-added scores.[45]

With the rapid proliferation of states and districts implementing or about to implement value-added models to evaluate teacher performance and use the results to inform high-stakes decisions, it is important that states and districts take a step back to evaluate the potential legal implications of doing so.

[45] CENTER FOR EDUCATOR COMPENSATION REFORM, DALLAS PRINCIPAL AND TEACHER INCENTIVE PAY PROGRAM, *available at* http://cecr.ed.gov/pdfs/profiles/Dallas.pdf.

PART III. CRITICISMS OF VALUE-ADDED MODELS

Many education researchers agree that in the majority of cases, value-added models are not reliable enough to determine whether a teacher should be fired.[46] For example, the Board on Testing and Assessment of the National Research Council of the National Academy of Sciences stated that value-added modeling "estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable."[47] In addition to the methodological concerns that make value-added measurements unreliable, the use of value-added models for high-stakes decision making also suffer from more basic concerns.

A. METHODOLOGICAL CONCERNS

1. NORM-REFERENCED RESULTS

First, it is important to understand that all value-added models report norm-referenced scores.[48] This means that all the teachers measured with a value-added model in a particular school, district, or state are compared to each other, and not to a criterion-based reference. Essentially, teachers are graded on a curve,[49] which means that every year, some teachers will be rated at the top, and other teachers will necessarily be rated at the bottom; "a district's logical aspiration to have exclusively 'high value-added' teachers is a technical impossibility."[50]

---

[46] *E.g.*, EPI BRIEFING PAPER, *supra* note 13.
[47] EPI BRIEFING PAPER, *supra* note 13, at 2.
[48] CORCORAN *supra* note 8, at 8.
[49] *Id.* at 9.
[50] *Id.*

Thus, when rating teachers, it matters a lot who the comparison group is. For example, assume that a value-added model was applied to all the teachers in *Best Elementary School*, widely considered to be the best elementary school in the *Best School District* in the state of *Worst Education in the Nation*. Upon receiving the value-added scores back, the principal of *Best School* learns that Ms. Smith has been rated the "worst" teacher at the school. However, because she teaches at the *Best Elementary School* in the *Best School District*, she is nowhere near the bottom of the *District's* distribution; in fact, she is at least an average teacher in the *District*. And when compared to all the other teachers in the state, she ranks in the top ten percent of all teachers. But since she teaches in the state of *Worst Education in the Nation*, the average teacher in other states still outperforms Ms. Smith, who, when compared with all the teachers in the country (if that were technically possible), may now appear to be only mediocre. It is therefore critical to know on what level retention decisions are being made (school, district, state, nation) and why.

2. POTENTIAL ERROR & INSTABILITY

Value-added models also suffer from potential error and instability. Potential error in value-added models comes from transitory influences that are outside of a teacher's control.[51] A measure with potential error is arbitrary because one year a teacher may receive a favorable score, and then next year an unfavorable one. In fact, research shows that at least ninety percent of the variation in student achievement gains is due to

[51] PETER Z. SCHOCHET & HANLEY S. CHIANG, U.S. DEP'T OF EDUC., ERROR RATES IN MEASURING TEACHER AND SCHOOL PERFORMANCE BASED ON STUDENT TEST SCORE GAINS 1 (2010) [hereinafter ERROR RATES]

factors outside of the teacher's control.[52] There are many potential sources of error that value-added models do not account for. While a few examples are provided here, the list is not exhaustive.

One potential source of error (and also bias, discussed below) is a student's gain or loss of learning over the summer months.[53] Typically, students from wealthier backgrounds have more opportunities to learn over the summer while students from poorer socioeconomic backgrounds are more likely to suffer, and to a greater degree, what is known as "summer learning loss."[54] Value-added models do not account for this, so teachers who teach students who grew over the summer will likely have a better value-added score than teachers who teach students who, during the summer months, forgot much of what they learned in the prior year.[55] Similarly, value-added models do not account for students who may perform worse in a given year because of something traumatic that happens in their life such as a divorce or a death in the family. Other potential sources of error include having a one or a few particularly disruptive students who impair instructional time in class one year, or experiencing subpar testing conditions, such as too much noise, on test day.[56]

Another important potential source of error is attributing student growth to the wrong teacher.[57] This could happen at the elementary school level, for example, with

---

[52] *Id.* at 35.
[53] EPI BRIEFING PAPER, *supra* note 13, at 15.
[54] *E.g.*, Karl L. Alexander, et al., *Lasting Consequences of the Summer Learning Gap*, 72 Am. Sociological Rev.167, 167-68 (2007); Douglas B. Downey, et al., *Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year*, 69 Am. Sociological Rev. 613, 613 (2004).
[55] EPI BRIEFING PAPER, *supra* note 13, at 12.
[56] CORCORAN *supra* note 8, at 18.
[57] EPI BRIEFING PAPER, *supra* note 13, at 12.

team teachers, or for teachers whose students may be pulled out of class for specialized instruction.[58] At the middle school and high school level the problem comes from the fact that a student has a number of teachers.[59] For instance, an excellent social studies teacher may make a student's performance in English better, even though the English teacher may not be very good. Similarly, a good chemistry teacher may teach skills that are important to a student's success in math. Since value-added models cannot account for these types of arrangements, one teacher will receive credit for work that should be attributed to other teachers.

Finally, many value-added models intentionally "fail to separate teachers' influence from the school's effect on achievement."[60] Performance across schools differs systematically because of differences in leadership, policy, staff quality, and the particular mix of students at a school.[61] This creates a problem because it confounds a teacher's effect with the school's effect.

Sources of potential error lead to instability in value-added models. Since two test scores are required to get a value-added measure, each teacher's score is affected by the measurement error not only in their own score, but the prior teachers' score as well, making the problem more complex.[62] "Given only one year of test score gains, it is impossible to distinguish between teacher effects and classroom-level factors."[63] One study that examined two consecutive years of data from five urban school districts

[58] *Id.* at 3.
[59] CORCORAN *supra* note 8, at 19.
[60] *Id.* at 18.
[61] *Id.*
[62] EPI BRIEFING PAPER, *supra* note 13, at 12.
[63] CORCORAN *supra* note 8, at 18.

showed that among teachers that ranked in the bottom twenty percent in effectiveness in the first year, less than a third remained in the bottom twenty percent the next year, and another third actually moved up into the top forty percent in teaching effectiveness.[64] Likewise, for teachers who were initially in the top twenty percent, only one third remained in the top twenty percent the second year, and another third moved all the way down into the bottom forty percent.[65] Many other studies have come to similar conclusions.[66]

Another study showed that "in a typical performance measurement system," more than twenty-five percent of teachers "who are truly average in performance will be erroneously identified for special treatments."[67] Similarly, more than twenty-five percent of teachers "who differ from average performance by 3 months of student learning in math or 4 months in reading" will be overlooked, and it is likely that these results are understated.[68] With ten years of data, it is estimated that the error rates will be reduced to twelve percent.[69] Moreover, "teachers' value added scores and rankings are most unstable at the upper and lower ends of the scale, where they are most likely to be used to allocate

[64] EPI BRIEFING PAPER, *supra* note 13, at 12.
[65] *Id.*
[66] *E.g.*, D. Aaronson, et al., *Teachers and Student Achievement in the Chicago Public High Schools*, 25 J. of Labor Econ. 95 (2007); D. Ballou, *Test Scaling and Value-Added Measurement*, 4 Educ. Finance and Policy 351 (2009); D. Goldhaber & M. Hansen, *Is it Just a Bad Class? Assessing the Stability of Measured Teacher Performance*, CRPE Working Paper No. 20085, Seattle, WA: Center on Reinventing Public Education (2008).
[67] ERROR RATES, *supra* note 50, at 35.
[68] *Id.*
[69] *Id.*

performance pay or to dismiss teachers believed to be ineffective."[70] Nevertheless, teachers persistently in the top or bottom five percent merit extra attention.[71]

Sources of error are magnified the smaller the number of students assigned to a teacher. "The larger the number of students in a tested group, the smaller will be the average error because positive errors will tend to cancel out negative errors."[72] Thus, the problem of instability will be particularly acute for teachers who teach small classes, such as elementary school teachers or special education teachers.[73] Increased instability is also more likely to occur for teachers who teach in schools with a highly mobile student population or in schools that experience high rates of absenteeism, since those students may miss the test.[74] Moreover, in order to have a value-added score, students need both a current-year score and a prior-year score, which exacerbates the problem for certain student populations. In the Houston Independent School District, for example, only 66 percent of students had both scores, "a fraction that falls to 62 percent for Black students, 47 percent for ESL [(English as a Second Language)] students, and 41 percent for recent immigrants."[75] To the extent that these students are not randomly distributed across the district, schools, and classrooms, value-added scores will also be biased.

There are two potential ways that value-added models can address the issues related to small class size, but each one presents a problem for using the results in high-stakes decisions. One approach involves just taking the scores as they are, without

[70] EPI BRIEFING PAPER, *supra* note 13, at 12.
[71] CORCORAN *supra* note 8, at 22-23.
[72] EPI BRIEFING PAPER, *supra* note 13, at 12.
[73] *Id.*
[74] *Id.*
[75] CORCORAN *supra* note 8, at 21.

making any statistical adjustments, but this results in too many teachers being classified as highly effective or highly ineffective.[76] The second approach, found in the random-effects model,[77] statistically "shrinks" the estimates of teachers with small classes toward the overall mean, which means their scores will be more similar to "the average effect of all teachers. This approach offsets the problem of distortions in the overall effects of teachers, but it makes identifying particularly effective or ineffective teachers who teach small classes considerably more difficult."[78]

Even if one were to assume that a teacher's movement from the top twenty percent to the bottom twenty percent were not due to error in the model, we would be left with the idea that the teacher moved down in the rankings due to other teachers improving their own value-added scores. This simply ties back to the fact that value-added models only provide relative rankings. For example, if Ms. Smith moved from the top twenty percent down to the bottom twenty percent, it might be that the District had reached its goal of improving the effectiveness of most of its teachers, but it does not mean that Ms. Smith is a bad teacher, not "adding value" to her students, or should be fired.

3. BIAS

Value-added models are also biased in important ways. A measure is biased when an unmeasured external influence causes systematic miscalculation. Bias is different than error, in that bias will always punish and reward teachers of the same types of students,

---

[76] RAND, THE PROMISE AND PERIL OF USING VALUE-ADDED MODELING TO MEASURE TEACHER EFFECTIVENESS 2 (2004) *available at* http://www.rand.org/pubs/research_briefs/RB9050.html.
[77] Daniel Koretz, *A Measured Approach: Value-Added Models Are a Promising Improvement, But No One Measure Can Evaluate Teacher Performance*, American Educator 18, 26 (Fall 2008).
[78] RAND, *supra* note 75, at 2.

whereas error may help a teacher's score one year and hurt it the next based on particular characteristics found in a particular class.

The bias in value-added models mainly stems from the fact that value-added models assume that students are randomly assorted across classrooms, schools, and districts, when, in fact, they are not. Within districts and states, students are frequently sorted based on their socioeconomic status, race, and/or nationality. Within schools, students rarely are randomly assigned to classrooms; students may be sorted into various classes based on academic tracking or based on teacher preferences, for instance.

Research shows that value-added models are correlated with the socioeconomic characteristics of the students; teachers who teach in high socioeconomic areas tend to have higher value-added scores than teachers who teach in low socioeconomic areas.[79] While the possibility that wealthier school districts are simply able to attract better teachers cannot be ruled out, one study examined the same teachers with different populations of students and found that the "teachers consistently appeared to be more effective when they taught more academically advanced students, fewer English language learners, and fewer low-income students."[80] For example, in 2010, the District of Columbia rated 663, or sixteen percent, of its teachers as "highly effective."[81] However, only five percent of those teachers came from the districts most impoverished area, while twenty-two percent came from the most prosperous.[82]

---

[79] *Id.*
[80] EPI BRIEFING PAPER, *supra* note 13, at 10.
[81] Bill Turque, D.C., *Teachers in Court Fight over Evaluations*, WASH. POST, July 1, 2011.
[82] *Id.*

Additionally, value-added models assume that students' rate of learning is the same regardless of the starting point of the student, even though there is no evidence to back up this assumption.[83] While it could be true that all students learn at equal rates, it could also be true that students with low academic achievement grow faster simply because they have more room to grow, or it could be that students with high academic achievement grow faster because "they have more knowledge and skills they can utilize to acquire additional knowledge and skills and, because they are independent learners, they may be able to learn as easily from ineffective teachers as from more effective ones."[84] Similarly, value-added models assume that a student's rate of learning is unaffected by their socioeconomic status.[85]

Because of the non-random sorting of students across states, districts, and within schools, it may be arbitrary (in the legal sense) to hold teachers accountable based on the model's results since the model's results are biased. Making retention and bonus decisions based on a biased model may also have legal implications involving disparate impact, but that is beyond the scope of this Article.

4. IMPRECISION

All statistical models are vulnerable to some level of uncertainty. Statistics use sample data to draw conclusions about the larger population. Because one sample will be different from another, each conclusion drawn will likewise be different. In value-added modeling, the sample used is the students in a teacher's classroom, which is then used to

[83] EPI BRIEFING PAPER, *supra* note 13, at 10.
[84] *Id.*
[85] *Id.*

draw a conclusion about a teacher's effectiveness. However, because a different set of students might lead to different conclusions about a teacher's effectiveness, the value-added model cannot be certain that any particular conclusion is correct.

Because of this uncertainty, the results of statistical models are generally presented as a range (or confidence interval). The width of a confidence interval that results from value-added modeling depends on the variance in student performance and the size of the sample, in this case, the number of students a teacher teaches. The less variance, and the larger the sample, the narrower the interval will be. Thus, confidence intervals will almost always be wider for teachers of students who are highly mobile or who are frequently absent because their sample sizes will be smaller as a result. However, there is little value-added models can do to alter the variance and sample size.

The level of confidence attached to the result can also affect the width of an interval. Most value-added models use a confidence level of ninety-five percent, which means that ninety-five percent of the time, the model gives the correct result; five percent of the time, it does not. It is impossible to know whether a particular result is one of the ninety-five percent that is correct. The higher the level of confidence used, the wider the interval. The level of confidence used in completely within the control of who designs the model. However, the trade-off in getting a narrower interval would be increasing the proportion of time that the results are incorrect.

Since value-added models rank teachers on a curve, the results are presented as a percentile ranking surrounded by a range (or confidence interval) of percentiles of where the teacher statistically could be. (Remember, five percent of the time, the range

presented will miss the teacher's "true" value-added.) While the given percentile ranking is a teacher's "most likely" estimate, the range must be accounted for as it "respresent[s] the extent of statistical precision with which the value-added estimate was calculated."[86]

Unfortunately, the confidence intervals that result from value-added models at a level of ninety-five percent certainty are quite wide. One study examined "the full set of value-added estimates reported to more than 12,700 teachers" in New York City.[87] When only one year of data was used, the average confidence interval for English teachers across the city spanned sixty-six percentile points.[88] That is, with one year of data, the value-added model could tell you, for example, that a teacher was between the first percentile and the sixty-sixth percentile. Thus, the teacher could be anywhere from the worst teacher in the district to well above average. One year of data for math teachers had confidence intervals that spanned sixty-one percentiles.[89] Of course, those numbers are only averages, so some teachers' confidence intervals would have had a narrower range, while some teachers would have had an even wider range.

Confidence intervals become narrower, however, when more years of data are included.[90] When three years of data were incorporated, the average confidence interval width for math teachers was reduced to thirty-four percentile points, and for English teachers was forty-four percentile points.[91] Still, however, the confidence intervals are quite wide.

---

[86] CORCORAN *supra* note 8, at 25.
[87] *Id.* at 23.
[88] *Id.*
[89] *Id.*
[90] *Id.* at 22.
[91] *Id.* at 23.

One problem with such wide intervals is that it makes it statistically impossible to distinguish teachers from one another. The results of this study showed that in math, half of teachers could not be distinguished from sixty percent of the math teachers in the same grade.[92] In English, three fourths of teachers could not be distinguished from sixty-three percent of all other English teachers.[93] Only about five percent of math teachers and three percent of English teachers "received precise enough percentile ranges to be distinguished from twenty percent or fewer other teachers."[94] Thus, using value-added models for high-stakes decisions can be problematic.

B. PROBLEMS WITH STANDARDIZED TESTS

Testing, in its current state, presents several potential problems for value-added models. First, in order for value-added scores to have any real meaning, tests must be able to reflect student growth. In order to reflect student growth, tests must be vertically aligned, have interval scaling, and be sufficiently difficult. The good news is, it is possible to implement well-designed tests, and perhaps the use of value-added growth models will lead to the implementation of tests that can better assess a student's ability to think critically. Currently, however, most standardized tests do not meet these requirements.

[92] CORCORAN *supra* note 8, at 25.
[93] *Id.*
[94] *Id.*

1. LACK OF VERTICAL & INTERVAL SCALING

Vertical scaling means that the test for each grade is linked to a common scale. If the tests are not linked to a common scale, then knowing that a student increased or decreased his or her test score by, say, fifteen points, does not mean much. Since tests are rarely vertically scaled, most value-added methods rescale the scores to have a mean of zero and a standard deviation of one.[95] This rescaling, however, introduces additional uncertainty into the results of value-added models.[96] One model, the percentile growth model, does not assume that tests are vertically scaled.[97]

Additionally, almost all value-added models assume that standardized exams have interval scaling, which refers to the idea that the difference between any two points on the scale means the same thing. In other words, it means that increasing your test score from a 10 to a 20 would reflect the same growth as moving from a 90 to 100.[98] Assuming interval scaling is problematic because students who start out at higher achievement levels may be growing more to get from the 90 to 100 than the student who grows from a 10 to 20, or vice versa.[99] Lack of interval scaling could lead to additional bias in value-added results.

---

[95] *Id.*
[96] HENRY I. BRAUN, USING STUDENT PROGRESS TO EVALUATE TEACHERS: A PRIMER ON VALUE-ADDED MODELS 14 (Educational Testing Services, 2005).
[97] DAMIAN W. BETEBENNER, A PRIMER ON STUDENT GROWTH PERCENTILES (National Center for the Improvement of Educational Assessment, 2008).
[98] Koretz, *supra* note 76, at 21.
[99] *Id.* at 26.

2. RANGE OF DIFFICULTY

Another problem with current standardized tests, which were designed to meet the mandates of No Child Left Behind and meeting minimum standards, is that they are likely too easy for high achieving students. How can you measure the growth of a student who scored a perfect score on the previous year's test? Even if that student scored a perfect score for the next teacher, it would reflect zero growth, and if the student happened to miss one or two questions, then it would reflect negative growth. On the flip side, if the test is too difficult overall, it would be difficult to show growth for low-achieving students. Thus, tests must present a range of questions to be sufficiently difficult for both low-achieving and high-achieving students to demonstrate growth.

3. VALIDITY

Tests should also be valid; they should measure the content that teachers are required to teach. While state curricula are often quite broad, standardized tests rarely reflect the entire curriculum. "Recent studies analyzing state test content in New York, Massachusetts, and Texas find that over many years of test administration, some parts of the state curriculum are never tested. To take one extreme case—the 2009 New York State eighth-grade math test—50 percent of the possible points were based on only seven of the forty-eight state standards…."[100] The degree to which tests reflect the curriculum typically varies across subjects and grades, depending on what content happens to be easily tested, which may now also have to be further narrowed into content that can be

[100] CORCORAN *supra* note 8, at 16.

vertically scaled more easily.[101] The problem with this is that teachers who teach the full curriculum might be penalized. Another possibility is that some teachers may be stronger at teaching a certain part of the content than other parts of the content. For example, an English teacher might be very good at teaching reading comprehension, but worse at sentence diagramming, or vice versa. Thus, "[g]iven two teachers of equal effectiveness, the teacher whose classroom instruction happens to be most closely aligned with the test—for whatever reason—will outperform the other in terms of value-added."[102]

Evidence shows that the choice of test (or what's on the test) can have dire consequences for teachers and value-added scores.[103] Houston ISD has administered two different standardized tests to its students at approximately the same time of year each year: the Texas Assessment of Knowledge and Skills (TAKS) and the nationally normed Stanford Achievement Test.[104] "[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." A good argument can therefore be made based on the narrowness of any given test in any given year, that what is selected out of the curriculum to be on the test in that particular year is arbitrary.

Still, a school district is likely entitled to implement the test of its choosing, so long as the test itself is a rational choice. Additionally, it seems perfectly reasonable to expect a teacher to be effective at teaching all parts of the curriculum, not just some parts.

---

[101] BRAUN, *supra* note 95, at 14.
[102] CORCORAN *supra* note 8, at 17.
[103] *Id.*
[104] *Id.*

Further, if the tests do switch between content material, then a teacher who was rated ineffective one year may be rated effective the next year as a result, which would preclude termination in most circumstances. Finally, if the tests really do only test the same narrow subject year after year, it should be easy enough for a teacher to figure that out and teach to that part of the test to improve his or her effectiveness score. Whether this behavior is desirable, however, is questionable.

Teaching to the test raises a larger question of to what extent the system of standardized testing for accountability purposes is really related to the stated goals of the education system. While in the past higher test scores generally have been correlated with higher income, evidence shows that "U.S. scores on international exams that assess more complex skills dropped from 2000 to 2006, even while state and local test scores were climbing, driven upward by the pressures of test-based accountability."[105] It will be interesting to see in the future how the correlation between standardized test scores and income holds up, given the increasingly intense focus on testing, the increasing narrowness of curricula tested on high-stakes tests in the United States, and the need for complex thinking skills in the global marketplace.

C. POLICY CONCERNS

1. MERIT-PAY IMPLEMENTATION & BUDGET CONCERNS

Rewarding especially effective teachers with additional pay sounds intuitively appealing. However, in addition to the fact that value-added results are unreliable, which

[105] EPI BRIEFING PAPER, *supra* note 13, at 7.

may reduce incentives to "improve" one's teaching, the structure of the bonus system and the size of the bonus system matter a lot in how teachers will respond. For example, if only the top five percent of teachers will be rewarded, and Ms. Jones repeatedly scores only near the fiftieth percentile and/or if she has students who are harder to "improve," then Ms. Jones may have little incentive to try to move up into the ninety-fifth percentile because she figures, no matter how hard she tries, she will never get there. So, where is her incentive to try any harder? This lack of incentive would be exacerbated if, like some teacher merit-pay plans, the bonus is rather small, such as $500.

On the other hand it is certainly possible to develop a merit pay plan that could properly incentivize more teachers to improve their teaching practice and their students' learning. More teachers could be reached, for example, by rewarding all of those who score average or better. While this may reward some teachers at the very top for no increased effort or growth, it will likely incentivize many more of those in the average and below-average effectiveness categories to improve. Another possibility would be to reward the entire school for average or better than average growth, as this would avoid the problems that may come with increasing competition between teachers in a school because it would incentivize collegiality and sharing among teachers, which has been shown to improve student outcomes.[106]

Still, given the seemingly constant refrain of schools, districts, and states facing "budget crises," which has become especially poignant in the last few years during the "Great Recession," there is little reason to believe that school districts will be able to

[106] *E.g.*, Roger Goddard, et al., *Collective Teacher Efficacy: Its Meaning, Measure, and Impact on Student Achievement*, 37 Am. Educ. Res. J. 479 (2000).

increase average teacher salaries significantly through bonuses since bonuses are "likely to come mostly from the redistribution of already-appropriated teacher compensation funds."[107] "If performance rewards do not raise average teacher salaries, the potential for them to improve the average effectiveness of recruited teachers is limited and will result only if the more talented of prospective teachers are more likely than the less talented to accept the risks that come with an uncertain salary."[108] While merit pay plans are intuitively appealing, it is difficult to see how they will be effective in improving the overall effectiveness of teachers without the injection of new funds and policies designed to incentivize below average and average teachers to improve their practice. However, with the infusion of new funds and a well-designed merit pay plan, improvement across the board may be possible.

2. SUPPLY OF TEACHERS

Value-added models also offer the promise of removing ineffective teachers from classrooms. However, in addition to evidence that shows it will be very difficult, if not impossible, to identify the least effective teachers (and that "least effective" is an inherently relative term), there is no evidence to support the notion that the teachers who are fired would be replaced by more effective teachers. [109]

However, that does not mean that school districts should continue to employ teachers who consistently rate poorly in effectiveness, especially when

[107] EPI BRIEFING PAPER, *supra* note 13, at 7.
[108] *Id.*
[109] *Id.* at 5.

those teachers have been given an opportunity to improve and have not. It would make the most sense then, in addition to removing the least effective teachers, to enact additional policy levers to ensure that people who will become more effective teachers are recruited into and retained by the profession.

D. PRINCIPALS' EVALUATIONS ARE INFLUENCED BY VALUE-ADDED RESULTS

So far, this paper has been concerned with the flaws in using value-added models to make high-stakes decision about teachers' jobs. However, no state or district has made value-added results the sole method on which to evaluate teachers; value-added models generally account for somewhere between 25-50 percent of a teacher's evaluation, although in Delaware, a teacher cannot be rated "effective" overall if the student growth expectations for the teacher's students are not met.[110] Other criteria generally include results from survey tools and principal observations.[111]

One study has shown that principals' subjective evaluations of their teachers were positively correlated with teacher value-added models.[112] This positive association should lend credibility to value-added assessments, or vice versa.

While a principal's observation or results from a survey may add credibility to a low value-added score, however, caution should be used in interpreting the results. First, as might be expected, the correlation between a principal's evaluation and a value-added evaluation was stronger when the value-added assessments had tighter confidence

[110] NCTQ, *supra* note 26, at 16.
[111] *Id.*
[112] Jonah E. Rockoff, et al., *Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools* (Nat'l Bureau of Econ. Research, Working Paper No. 16240, 2011).

intervals.[113] Experienced principals were also better at predicting a teacher's value-added than newer principals.[114]

This same study also showed that when principals received value-added information about their teachers, they changed their own subjective beliefs about those teachers' productivity.[115] While this reaction to new information is to be expected, it should be troubling from a teacher's perspective because it essentially means that the value-added score is being double counted in their evaluation: first from the score itself, and then from the influence the score has on a principal's beliefs about the teacher's effectiveness, which will appear now in the principal's evaluation of the teacher. To the extent that the value-added prediction provided an incorrect assessment of a teacher's effectiveness, the principal's evaluation may now exacerbate the mistake, rather than mitigate it.

While no research has been conducted on whether parental surveys would be similarly influenced by value-added results, the possibility remains, especially in areas where value-added scores are released publicly, as has been the case in Los Angeles and New York City. The results could be particularly stark with parental surveys since the principals in the study discussed above went through a three-hour training on how value-added models work, whereas parents would not be similarly trained.

[113] *Id.* at 20.
[114] *Id.* at 21.
[115] *Id.*

PART IV. POTENTIAL SUBSTANTIVE DUE PROCESS IMPLICATIONS OF USING VALUE-ADDED MODELING IN HIGH-STAKES DECISIONS ABOUT TEACHERS

With many expert criticisms of value-added models, it is possible for some districts to face substantive due process challenges from teachers who were negatively affected based on their value-added results. This section considers the extent to which substantive due process challenges are likely to succeed.

A. DUE PROCESS

There are two types of due process guaranteed under the U.S. Constitution: substantive due process and procedural due process. Specifically, these rights are provided by the Fifth and Fourteenth Amendments, which prohibit the federal and state governments, respectively, from depriving any person "of life, liberty, or property without due process of law."

Substantive due process limits the government's ability to take away a person's life, liberty, or property by requiring the government to have a sufficient justification for doing so. What is considered a "sufficient justification" depends on the substantive right(s) involved. Generally, substantive rights are categorized into two levels: fundamental rights and all other rights.

Fundamental rights, such as the right to vote, the right to privacy, and the right to freedom of speech, trigger what is known as "strict scrutiny." Strict scrutiny requires that the government law or action be narrowly tailored to further a compelling government interest. This requires the government to prove that the law is the least restrictive

alternative available. The application of strict scrutiny is so hard for the government to overcome that is has been referred to as "strict in theory and fatal in fact."[116]

For all other substantive rights, the "rational basis test" is used, which means the government law or action will be upheld if it is rationally related to a legitimate government purpose.[117] The rational basis test requires the person challenging the government action to prove that the law has no "conceivable legitimate purpose or that it is not a reasonable way to attain the end," [118] or that it is "arbitrary and capricious, does not achieve or even frustrates a legitimate state interest, or is fundamentally unfair."[119] Relatively few cases have been successful on this front.

Procedural due process refers to the procedures the government must follow before depriving a person of life, liberty, or property. Generally, this refers to the type of notice and hearing the government must provide to an individual or group whose substantive rights are at stake. If no substantive rights are at stake, then procedural due process is not required.[120] Thus, once an employee establishes a substantive right in their employment, the employee is entitled to procedural due process as guaranteed by the Constitution and may also attempt to prove that the government's basis for the employee's termination was irrational.

---

[116] Gerald Gunther, *Foreword: In Search of Evolving Doctrine on a Changing Court: A Model for a Newer Equal Protection*, 86 Harv. L. Rev. 1, 8 (1972).
[117] *See, e.g.*, Williamson v. Lee Optical, 348 U.S. 483 (1955); Day-Brite Lighting, Inc. v. Missouri, 342 U.S. 421 (1952).
[118] ERWIN CHEMERINSKY, CONSTITUTIONAL LAW: PRINCIPLES AND POLICIES 540 (3d edition, Aspen, 2006).
[119] Debra P. v. Turlington, 644 F.2d 397, 404 (5th Cir. 1981).
[120] *E.g.*, Town of Castle Rock v. Gonzales, 545 U.S. 748, 756 (2005) ("The procedural component of the Due Process Clause does not protect everything that might be described as a 'benefit' ….").

B. DO TEACHERS HAVE A PROPERTY INTEREST IN THEIR JOBS?

The Supreme Court has struggled with the question of whether government employment is a property right or merely a privilege. Initially, the Supreme Court considered government jobs a privilege, meaning an employee was not entitled to due process rights upon termination.[121] In 1972, however, the Supreme Court explicitly discarded the "rights-privileges" framework for determining when due process attached.[122] The Supreme Court has stated that in order "[t]o have a property interest in a benefit, a person clearly must have more than an abstract need or desire for it. He must have more than a unilateral expectation of it. He must, instead, have a legitimate claim of entitlement to it. It is a purpose of the ancient institution of property to protect those claims upon which people rely in their daily lives, reliance that must not be arbitrarily undermined."[123] Thus, property rights were dependent on how important the benefit was to an individual. At the same time, however, the Court stated that property rights are only "created and their dimensions…defined by existing rules or understandings that stem from an independent source such as state law…."[124] Thus, a state can deny a person's property right in government employment simply by codifying whether a property right exists, and to what extent. It is this latter pronouncement that typically governs whether a government employee has a property interest in his or her job. Thus, despite its explicit

[121] *See, e.g.*, Bailey v. Richardson, 182 F.2d 46 (D.C.C. 1950), *affd. by an equally divided Court*, 341 U.S. 918 (1951).
[122] Bd. of Regents v. Roth, 408 U.S. 564, 571 (1972).
[123] *Id.* at 577.
[124] *Id.*

decree, the Supreme Court appears to have never totally disregarded the "rights-privileges" distinction for government employment.

Generally, in order to have a property right in their jobs, teachers must have a reasonable expectation of continued employment.[125] In *Roth*, a teacher at Wisconsin State University was hired for a fixed term of one academic year, and he was timely informed that he would not be rehired at the end of that term.[126] The teacher's contract provided that "no reason for non-retention need be given. No review or appeal is provided in such case," and none were provided. [127] Still, the teacher argued that the University's failure to give him "any reason for non-retention and an opportunity for a hearing violated his right to procedural due process of law."[128] The Supreme Court disagreed because based on his contractual terms, the teacher had no reasonable expectation of being rehired, and thus no property interest inhered.

A formal, contractual tenure system, however, is not required to create an expectation of continued employment (although if one is in place, a property interest clearly attaches).[129] For four years, under a series of one-year contracts, Robert Sindermann was a professor in the state college system in the State of Texas.[130] In his fourth year of teaching, Sindermann became involved in public disagreements with the policies of the college's Board of Regents, and at the end of that year, the Board of

[125] Bd. of Regents v. Roth, 408 U.S. 564 (1972); Perry v. Sindermann, 408 U.S. 593 (1972).
[126] Bd. of Regents v. Roth, 408 U.S. 564, 567-68 (1972).
[127] *Id.* at 567.
[128] *Id.* at 569.
[129] Perry v. Sindermann, 408 U.S. 593 (1972).
[130] *Id.* at 594.

Regents terminated his employment.[131] Like the teacher in *Roth*, Sindermann argued that the college's failure to provide him an opportunity for a hearing violated his right to procedural due process.[132] Unlike the teacher in *Roth*, however, Sindermann was able to demonstrate a reasonable expectation of continued employment due to the college's "de facto tenure program," which stemmed from a provision in the college's official Faculty guide that stated that despite the lack of a formal tenure system, "[t]he Administration of the College wishes the faculty member to feel that he has permanent tenure as long as his teaching services are satisfactory…."[133] Thus, a teacher must be "given an opportunity to prove the legitimacy of his claim" to a property interest in his job "in light of the policies and practices of the institution."[134] While proof of a property interest does not guarantee reinstatement, it does obligate an educational institution to provide a hearing where the teacher is informed of the grounds for non-retention and is allowed to challenge their sufficiency.[135]

For public K-12 teachers, expectations of a property interest in their jobs generally are dependent on state statute. Teachers who are only on probationary contracts do not have a property interest in their jobs if their contracts are not renewed at the end of the contractual period. A school district may choose not to renew a contract for any reason, or no reason at all, unless the reason infringes on a teacher's right to participate in a constitutionally protected activity, or if the school publicly attached a stigma to the

[131] *Id.* at 595.
[132] *Id.* at 595.
[133] *Id.* at 600.
[134] *Id.* at 603.
[135] Perry v. Sindermann, 408 U.S. at 603.

teacher as a reason for termination.[136] However, if a school district terminated a teacher's probationary contract in the middle of the contract, then the teacher has a property interest in continuing their job and would thus be entitled to procedural due process.

Many states have laws in place that grant teachers tenure or continuing contracts. To earn tenure in Virginia, for example, a teacher must teach a probationary term of three years in the same school division.[137] Once tenure is earned, teachers gain a property interest in their jobs and may only be fired for cause. Most state tenure statutes identify specific causes for dismissal, such as incompetence, insubordination, and/or immorality, or they may provide that teachers can be fired for any "good cause."[138] Thus, whether an employee has a property right in a job depends upon whether there is a statutorily created right to one. Otherwise, the job is merely a privilege.

With the many new laws tying a teacher's employment and/or tenure to a value-added score, an important question arises: Do these laws give teachers a property interest in their jobs? In some cases, the answer is clearly "no." Florida, for example, has eliminated teacher tenure altogether.[139] Thus, teachers in Florida no longer have a property interest in their jobs.

Other states have retained traditional teacher tenure. New York, for example, still allows teachers to earn tenure, but has added that teachers with two consecutive annual "ineffective" ratings may be charged with incompetence through an expedited hearing

[136] Scheelhaase v. Woodbury Central Community Sch. Dist., 488 F.2d 237, 242 (8th Cir. 1973); LOUIS FISCHER, DAVID SCHIMMEL, & LESLIE R. STELLMAN, TEACHERS AND THE LAW 233 (7th ed.2007).
[137] VA. CODE ANN. § 22.1-303 (Westlaw 2012).
[138] SABA BIREDA, DEVIL IN THE DETAILS: AN ANALYSIS OF STATE TEACHER DISMISSAL LAWS 6 (Center for American Progress 2010).
[139] NCTQ, *supra* note 26, at 25.

process.[140] Thus, in New York, it is clear that tenured teachers retain a property interest in their jobs, and thus substantive due process attaches.

The laws in Indiana, however, provide a trickier situation. While Indiana has retained teacher tenure, the law provides that a teacher's tenure can be revoked, and the teacher returned to probationary status, if the teacher has received an "ineffective" rating.[141] Teachers who fail to improve after being returned to probationary status are subject to termination.[142] Thus, it appears that for a teacher who receives an "ineffective" rating, rather than terminate the teacher on that basis, which would require a hearing in which the state would have to defend its decision, the state simply revokes the teacher's tenure in their job, and then after a year can terminate the teacher for any reason, or no reason at all.[143] The reasoning behind this structure is unclear, and it raises the question regarding the extent to which the government can define when a property interest exists.

C. ARE VALUE-ADDED MODELS ARBITRARY AND CAPRICIOUS?

For states like New York, where teachers are tenured, or for states where it is questionable whether teachers enjoy a property right in their jobs, the following analysis is intended to shed light on the extent to which value-added models are "rational" in a legal sense (i.e., would not lead to arbitrary and capricious teacher terminations). For states that employ teachers at-will, teachers have no property interest in their job, and the following analysis is therefore irrelevant for legal purposes, although it should still give

[140] *Id.* at 16.
[141] *Id.* at 24.
[142] *Id.*
[143] *See id.*

policy-makers pause when considering using value-added models for high-stakes decisions.

Teachers who have a property interest in their jobs who are fired on the basis of a value-added model will be required to prove either that the law has no rational basis or that their termination based on the model was arbitrary. Under the rational basis test, a law will be upheld if it is rationally related to a legitimate government purpose.[144] Thus, a teacher must show that the teacher evaluation system has no "conceivable legitimate purpose or that it is not a reasonable way to attain the end."[145]

The rational basis test is a very difficult bar to overcome. Only a few cases have been successful on this front, mostly on the basis of violating the Equal Protection Clause of the Fourteenth Amendment, not the Due Process Clause.[146] Additionally, a law will be upheld if any conceivable legitimate purpose is served, not simply the purpose that the lawmakers actually intended.[147] Added to all of this is the fact that courts are traditionally very deferential to decisions made in an educational context, including decisions regarding teacher competency: "For sound policy reasons, courts are loathe [*sic*] to intrude upon the internal affairs of local school authorities in such matters as teacher competency."[148] These combined factors make the possibility of overcoming the rational basis test remote.

---

[144] *See, e.g.* Williamson v. Lee Optical, 348 U.S. 483 (1955).
[145] CHEMERINSKY, *supra* note 117 at 540.
[146] *See, e.g.*, Romer v. Evans, 517 U.S. 620 (1996); City of Cleburne v. Cleburne Living Center, Inc., 473 U.S. 432 (1985).
[147] McGowan v. Md., 366 U.S. 420, 426 (1961).
[148] Blunt v. Marion County School Board, 515 F.2d 951, 956 (5th Cir. 1975).

Nevertheless, the Supreme Court has been criticized for unevenly applying the rational basis test; sometimes it employs what is known as rational basis "with bite," and other times "the application of traditional rational basis review in one case is often quite distinct from its iteration in other cases."[149] Additionally, the law presumes certified teachers to be competent, and the burden rests with the school board to prove incompetency.[150] One court has stated that a "teacher's dismissal is arbitrary and capricious if he can prove that each of the stated reasons underlying his dismissal is trivial…or is wholly unsupported by a basis in fact."[151] (Many courts or states, however, require the school board to satisfy the "substantial evidence" standard.[152] The test for substantial evidence is generally viewed by courts as requiring greater judicial scrutiny than does the arbitrariness standard.[153] The test for whether the substantial evidence test has been satisfied is whether, based on the evidence, "reasonable minds could have reached the same conclusion."[154]) Consequently, there are cases where individual teachers have been able to show their terminations were arbitrary and capricious.[155] Therefore, it remains possible for a teacher to prove that his or her termination based on a value-added model was legally improper.

---

[149] *E.g.*, Neelum J. Wadhwani, Note, *Rational Reviews, Irrational Results*, 84 Tex. L. Rev. 801, 801 (2006).
[150] Gene S. Jacobsen, *The Dismissal and Non-Reemployment of Teachers*, 1 J.L. & Educ. 435, 437 (1972).
[151] Fisher v. Snyder, 476 F.2d 375, 377 (8th Cir. 1973).
[152] *E.g.*, Lee v. Tuscaloosa County Bd. Of Educ., 591 F.2d 324 (1979); Miller v. Houston Indep. Sch. Dist., 51 S.W.3d 676 (Tex. App.—Houston (1st Dist.) 2001).
[153] Charles Alan Wright & Charles H. Koch, Jr., *Substantial Evidence or Reasonableness*, 33 Fed. Prac. & Proc. § 8333 (1st ed.).
[154] *E.g.*, Miller v. Houston Indep. Sch. Dist., 51 S.W.3d at 680 (Tex. App.—Houston (1st Dist.) 2001).
[155] *E.g.*, Trustees, Missoula County Sch. Dist. No. 1. v. Anderson, 757 P.2d 1315 (Mont. 1988); Collins v. Faith Sch. Dist. No. 46-2, 574 N.W.2d 889 (S.D. 1998).

1. CHALLENGING THE SYSTEM OF VALUE-ADDED MODELS

The underlying purpose of evaluating teachers using a value-added model is clearly rational: States and school districts want to identify and retain effective teachers, identify the best teachers to help identify best practices, and remove at least the very worst teachers from the classroom in order to improve student achievement. Since research demonstrates that teacher quality is the single most important school-related factor that bears on student achievement,[156] this makes sense. Being able to identify the characteristics of effective teachers would be a boon for education. For example, educator preparation programs would know what skills to teach to aspiring teachers, and school districts could potentially know who to hire (and not hire) at the earliest stages of the hiring process.

Since the purpose of using value-added models is rational, the main barrier for a teacher who was fired on the basis of one would be to demonstrate that using value-added models at this point in their development in high-stakes decision "is not a reasonable way to attain the end,"[157] or that the system, when used in high-stakes decisions is "arbitrary and capricious, does not achieve or even frustrates a legitimate state interest, or is fundamentally unfair."[158]

It is unlikely that a court could be convinced that the whole system of using value-added modeling to evaluate teachers for high-stakes decisions is arbitrary and capricious. Prior to value-added models, teacher evaluations were

---

[156] JENNIFER KING RICE, UNDERSTANDING THE EFFECTIVENESS OF TEACHER ATTRIBUTES, v (Econ. Policy Institute 2003).
[157] CHEMERINSKY, *supra* note 117 at 540.
[158] Debra P. v. Turlington, 644 F.2d 397, 404 (5th Cir. 1981).

primarily conducted by subjective evaluations based on objective criteria. Courts have previously upheld teacher terminations based on these types of evaluations, even though they were more subjective, and therefore potentially more arbitrary, than value-added models.

Courts have upheld, for example, the firing of teachers based on evaluations that were purely subjective in nature, even where the evaluations provided conflicting conclusions. In several cases, teachers have been fired on the basis of the observations of several principals, some of whom rated the teacher unfavorably, and some of whom recommended the teacher be retained.[159] Thus, a system comprised only of subjective principal evaluations in which the principals disagreed was a rational way to dismiss a teacher.

In another case, however, the court did invalidate a teacher's evaluation as arbitrary and capricious. In *Scheelhaase*, a teacher was fired for incompetence because of "the low scholastic accomplishment of her students on Iowa Tests of Basic Skills and Iowa Test of Educational development."[160] The District Court held for the teacher because "professional competence cannot be determined solely on the basis of her students' achievement on the ITBS and ITED, especially where students maintain normal educational growth rates."[161] In other words, basing a teacher's evaluation based on raw student performance is irrational

[159] Fox v. San Francisco Unified Sch. Dist., 245 P.2d 603, 607 (Cal. Ct. App. 1952); In re Proposed Termination of James E. Johnson's Teaching Contract with Indep. Sch. Dist. No. 709, 451 N.W.2d 343 (Minn. Ct. App. 1990).
[160] Scheelhaase v. Woodbury Central Community Sch. Dist., 349 F.Supp. 988, 989 (N.D. Iowa 1972), *rev'd on other grounds*, 488 F.2d 237 (8th Cir. 1973).
[161] *Id.* at 990.

because it fails to show the effect a teacher had on student learning; it is mostly a reflection of how much students knew prior to entering the class.

While this case shows that a court may invalidate a method of evaluation, it also demonstrates that what was most important was that students were learning at a normal rate; a teacher should not be penalized for teaching underachieving students so long as the teacher is supporting at least normal student learning growth rates, which is exactly what value-added models measure, and in an objective way. Further, when implemented properly, value-added models can provide valid results. Because of this, and the lenient rational basis standard, courts will almost certainly validate the use of value-added modeling as a whole.

2. Challenging an Individual Teacher Termination

Still, a teacher may have room to prove that his or her individual termination based on a value-added score was arbitrary and capricious. This is because an objective result may be more amenable to judicial scrutiny. Courts are loath to question an administrator's judgment regarding teacher incompetency because administrators must use professional judgment, which is inherently subjective:

> A teacher's competence and qualifications for tenure or promotion are by their very nature matters calling for highly subjective determinations, determinations which do not lend themselves to precise qualifications and are not susceptible to mechanical measurement or the use of standardized tests. These determinations are "in an area in which school officials must remain free to exercise their judgment"…Courts are not qualified to

> review and substitute their judgment for these subjective, discretionary judgments of professional experts.[162]

This implies that courts may be in a better position to scrutinize an evaluation that is more objective. With value-added models, teacher competence is now "susceptible to mechanical measurement."[163] Thus, the idea of subjecting the results of value-added models to judicial scrutiny, especially when it is known that the results are vulnerable to error, may be more amenable to courts. In fact, administrative courts, for example, regularly deal with complex statistical materials when judging a government agency's particular course of action. Thus, courts may be more willing to review the result of a value-added model, or subject it to more intense criticism because precisely because it is objective. Because the models' shortcomings are well known, it may in certain cases lead to results that are objectively wrong and therefore arbitrary. Thus, while the objectivity of the models may make them overall less arbitrary than wholly subjective assessments, the objectivity offered by value-added models may also make them more vulnerable to court review.

Still, at least one court is of the opinion that because "[t]here are few, if any, objective criteria for evaluating teacher performance….[e]ach case must, therefore, be assessed on its own facts."[164] This suggests, counter to the argument

[162] Clark v. Whiting, 607 F.2d 634, 639-640 (4th Cir. 1979).
[163] *Id.*
[164] Hollingsworth v. Bd. of Educ. of Sch. Dist. of Alliance, 303 N.W.2d 506, 508 (Neb. 1981) (citing Sanders v. Bd. Of Educ., 263 N.W.2d 461 (Neb. 1978)).

above, that a court may in fact be less inclined to assess each case on its own facts if an objective measure of ineffectiveness is provided.

Nevertheless, because no teacher evaluation is based solely on value-added measures and because value-added models contain several potentially critical weaknesses, courts will likely review the value-added data in concert with other factors that have traditionally informed teacher evaluations. If everything on the evaluation is negative, it will be easy for a court to dismiss a teacher's case. However, where information conflicts, courts will have to examine all of the evidence, which will give the teacher the opportunity to challenge either the value-added results, or the principal's subjective evaluation, depending on which part of the evaluation was negative.

In some cases, however, teachers may not even need to present subjective evaluations that conflict with the results of a value-added score. In many cases value-added results for an individual teacher may simply be inaccurate (e.g., too much bias in the model, wildly disparate results from year to year). In other cases, value-added models may be implemented in such a way that would lead to arbitrary results (e.g., terminating a teacher on the basis of a single year's score; a teacher is linked to the wrong students). Therefore, it is entirely reasonable that some teachers may be able to prove their terminations were arbitrary and capricious.

Although there are many potential factual permutations that would give an individual teacher a good case for challenging a dismissal based on the results of a

value-added score, I will focus on just one illustrative example based on actual events that highlights several of the potential problems in the use and implementation of value-added models.

Sara Wysocki was a new teacher in Washington, D.C., where value-added scores make up fifty percent of a teacher's evaluation.[165] D.C.'s evaluation system, IMPACT, grades teachers on a 1-4 scale: ineffective, minimally effective, effective, and highly effective.[166] Wysocki taught fifth grade at MacFarland Middle School, where eighty percent of students qualify for free-or-reduced lunch, and less than thirty percent scored proficient on the city's 2010 reading test.[167] In her first year of teaching, she was rated "just below effective" based on classroom observations, and her value-added score was low, leaving her overall rating at "minimally effective."[168] By her second year, her classroom observations scores improved to 3s and 4s, and administrators were praising her for her involvement with students and innovative ways of engaging parents.[169] However, fourteen of her twenty-five students attended fourth grade at Barnard Elementary, which was found to have "unusually high numbers of answer sheet erasures in spring 2010, with wrong answers changed to right.[170] Twenty-nine percent of Barnard's 2010 fourth graders scored at the advanced level in reading,

[165] Bill Turque, '*Creative…Motivating' and Fired*, WASH. POST, Mar. 6, 2012.
[166] *Id.*
[167] *Id.*
[168] *Id.*
[169] *Id.*
[170] *Id.*

about five times the District average."[171] Wysocki's students scored below their "predicted average," which meant her value-added scores were low, and brought her overall rating down to "minimally effective" for the second year in a row. As a result, Wysocki was fired.

While Wysocki's individual score may "not reflect a deeper issue with IMPACT,"[172] it certainly raises questions about whether she was fired arbitrarily. There are several points worth noting about the potential arbitrariness of Wysocki's termination. First, she was a brand new teacher (although she had previously been a teaching assistant in a private Waldorf school,[173] a type of school that de-emphasizes testing and focuses on creativity, emotions, aesthetics, social sensitivity, willpower, and people's moral nature).[174] New teachers are commonly believed, even expected, to be inferior to more experienced teachers, due to a steep learning curve.[175] It is arguable, therefore, that subjecting new teachers to high-stakes decisions based on value-added models is irrational because they will be forced out of the profession before they even have time to learn it.

Second, Wysocki taught in a high-poverty school, which could suggest her scores at least partially reflect a bias in the model. Third, the possibility that her students cheated was a real one, but the hearing panel wrote that "[t]he Board and

---

[171] Turque, *supra* note 165.
[172] *Id.*
[173] *Id.*
[174] Association of Waldorf Schools of North America, Waldorf Education: Frequently Asked Questions *available at* http://www.whywaldorfworks.org/02_W_Education/faq_about.asp.
[175] D. Boyd, et al., *supra* note 3, at 177.

the Chancellor note that investigations of cheating are outside the scope of the Chancellor's appeals process."[176] If cheating had previously occurred, it would not only provide substantive support for the claim that Wysocki's termination was arbitrary, but also calls into question the sufficiency of the procedural process afforded to her. Finally, while wide disparities between classroom observations and value-added scores are "quite rare,"[177] they clearly occur, and those teachers should warrant special attention. Giving such teachers' ratings extra attention would not be too onerous for a district since it happens so rarely, and it could save the District unnecessary litigation. If disparities in evaluations frequently occur, then the model, the training of the principal(s), or both need to be investigated, or it could call the entire system into question.

A teacher might also have a good case where his or her scores were truly unstable—where a teacher may have been rated effective one year, and then ineffective for one year, then average, and then ineffective. While it may not be entirely arbitrary to fire a teacher such as this, a court may balance a teacher's substantive right to his or her job with the district's desire to fire ineffective teachers on the basis of a value-added model that produces unstable results. Thus, the teacher should at least have the opportunity to demonstrate that he or she is effective notwithstanding the value-added scores and/or that the results are flawed and should therefore not be accorded serious weight in his or her particular case.

[176] Turque, *supra* note 165.
[177] *Id.*

While a court may be more willing to review the criticisms of value-added modeling, or poor implementation of the value-added model, it still remains unlikely that the whole system would be rejected because its underlying purpose is rational and when implemented properly, the models can produce valid results. Thus, it seems that if a claim against using value-added models were to succeed, it would be because an individual teacher was arbitrarily let go because of the particular facts that led to a low value-added score.

D. TEACHER REMEDIATION

Another potential area where a teacher termination could be held arbitrary and capricious occurs at the intersection of teacher remediation and value-added models. Allowing teachers the opportunity for remediation is not required unless it is stated in a state statute or district policy that teachers are entitled to it. Several states, including Colorado and Indiana, have given teachers who score poorly on value-added models a period for remediation. It is unclear, however, how a teacher can be remediated successfully since a value-added score does not tell a teacher what is wrong with his or her teaching practice or how to improve. At best, it seems, if the data are disaggregated, the teacher may find a particular area of content that could be improved. This fact raises important questions.

Previously, teachers received remediation based on observations of how the teacher was teaching or interacting with students—all inputs that the teacher could attempt to change and principals could readily observe the change to determine whether the teacher had "improved." This appears to be the first time, however, that a teacher

will receive a remediation plan due to an objectively poor output—the value-added score—but will be offered a plan to improve that output based necessarily on observation of the teacher's behavior. An interesting legal question will be posed if and when a teacher faithfully follows the remediation plan provided by the evaluator, and the evaluator is satisfied through his or her observations that the teacher has "improved," but the teacher fails to improve his or her value-added score. Is the evaluator at fault for suggesting remediating things that do not lead to improved learning? Is it arbitrary to fire a teacher who failed to improve despite implementing the remediation plan, or is it further proof that the teacher is incompetent, and simply cannot be remediated?

PART V. RECOMMENDATIONS & CONCLUSION

States and districts should be wary of implementing value-added models for high-stakes decisions because methodological problems with value-added models, the limitations of current standardized tests, and other policy problems demonstrate that value-added models are currently flawed. However, whether a teacher will be able to demonstrate that the system as a whole is "not a reasonable way to attain the end" is questionable. Value-added models clearly have a rational goal and currently offer the only way to objectively isolate and measure a teacher's contribution to student learning. For a court to throw out the entire system, the implementation of the value-added model would have to be so bad that it would pervert the entire purpose of using it in the first place. This is unlikely to be shown.

Individual teachers, however, may realistically be able to show that their termination based on a value-added score was arbitrary and capricious, although it will still be a difficult hurdle. Teachers who score persistently in the bottom few percentiles (assuming that the district is not filled entirely with exceptionally effective teachers), and who also consistently receive poor feedback on other evaluation criteria, will probably never be able to meet their burden of proof. Other teachers, as demonstrated by the Wysocki story, have a much more egregious case to present. With all the potential permutations in value-added models, implementations of value-added models, and the individual cases of

teachers who may be fired as a result, many will be able to demonstrate that their terminations were arbitrary.

To avoid adverse legal decisions, then, school districts and states should use caution in selecting a value-added model and deciding how to implement it. There are many steps a state or school district could take to obviate substantive due process claims against them.

From a purely legal perspective, to avoid substantive due process claims altogether, states and/or school districts should eliminate tenure. Eliminating tenure, however, may not be politically feasible and it may, for various reasons, not be desirable from a policy perspective. For example, using a system that appears arbitrary, even if it does not meet the legal definition of "arbitrary," could lower teacher morale across the board and/or it could lead to teachers having less incentive to improve their teaching, since their scores are perceived as being unrelated to what they are doing in the classroom.

In states or districts where teachers have tenure, retention decisions should never be made on the basis of a single year's test scores. Depending on the type of model, two years of value-added scores may be sufficient, if they are consistent and extremely low (i.e., below the fifth percentile), and if the teacher can be sufficiently distinguished from other teachers. Generally, however, a better practice would be to base termination decisions on three years of data, due to the instability of value-added models and wide confidence intervals in the results from value-added models.

To mitigate the problems that can result from summer learning loss, school districts could use benchmark testing at the beginning of the school year, which would help hold teachers accountable only for learning that takes place during the school year, and not learning (or loss of learning) that takes place over the summer, which is out of the teacher's control. In this way, teachers who teach students from high socioeconomic backgrounds who are more likely to continue their learning over the summer will be less likely to have an unfairly high value-added score, while teachers who have students from low socioeconomic backgrounds who are more likely to experience learning losses over the summer will not be penalized for teaching economically disadvantaged students.

Districts should keep in mind, however, that student performance on tests may vary based on the stakes attached to the test. If benchmark tests affect a student differently than the end-of-year test, student performance on each test can be expected to vary accordingly. Thus, if benchmark tests have no effect on a student's grade, for example, while the end-of-year test is tied to grade promotion, a student is likely to try much harder on the end-of-year test than on the benchmark test, which could lead to a value-added score higher than it would be otherwise. Nevertheless, it may be impossible to tie the same consequences to each test. Benchmark testing may also provide an incentive for teachers to "game the system," as teachers could persuade students to perform less than their true ability on a beginning of the year test, which would help inflate a teacher's value-added score.

For any value-added model except for percentile growth models, states should ensure that standardized exams are vertically scaled so that student growth can be more

accurately measured. This will prevent a certain grade-level of teachers from reaching unfair gains if their grade's test was relatively easier than the prior grade's test, and vice versa.

Rather than fire new teachers based on their value-added scores, allow them sufficient time to improve. If new teachers do not show sufficient improvement after three or four years, then termination would be warranted.

Certainly, however, new teachers whose value-added results are below "effective" should not be awarded tenure based on length of service. Rather, tenure should now be tied to having a sufficient number of years of "effective" value-added scores. That way, at the end of several years, if a teacher is still performing poorly, it will be easier for a district to dismiss the teacher.

Finally, districts should develop a process to further investigate situations in which a teacher's classroom evaluation markedly differs from the teacher's value-added results. This will help ensure a district provides sufficient procedural due process, and a basic sense of fairness, to a teacher whose value-added scores do not line up with other the results from other types of evaluation.

While no evaluation system is perfect, and value-added models offer potentially enormous benefits to school districts based on objective criteria, states and districts should be careful to implement value-added models and any corresponding high-stakes decisions appropriately. Doing so will help balance the tension between a district's desire to identify and remove ineffective teachers (and reward the most effective ones) with a teacher's desire to be fairly evaluated and right not to be arbitrarily dismissed.

WORKS CITED

American Recovery and Reinvestment Act of 2009, Pub.L. No. 11-5 (H.R.1), 123 Stat. 115 (Feb. 17, 2009) as amended by Pub.L. No. 111-8 (H.R. 1105), the Omnibus Appropriations Act, 2009, Division A, Section 523, 123 Stat. 524 (Mar. 11, 2009)." 2009.

Association of Waldor Schools of North America. *Waldorf Education: Frequently Asked Questions.* n.d. http://www.whywaldorfworks.org/02_W_Education/faq_about.asp (accessed April 25, 2012).

B. Rown, R. Correnti, R.J. Miller. "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools." *Teachers College Record* 104 (2002): 1525.

Bailey v. Richardson. 182 F.2d 46 (D.C. Circuit, 1950).

Ballou, D. "Test Scaling and Value-Added Measurement." *Education Finance and Policy* 4 (2009): 351.

Betebenner, Damian W. *A Primer on Student Growth Percentiles.* National Center for the Improvement of Educational Assessment, 2008.

Bireda, Saba. *Devil in the Details: An Analysis of State Teacher Dismissal Laws.* Center for American Progress, 2010.

Blunt v. Marion County School Board. 515 F.2d 951 (5th Circuit, 1975).

Board of Regents v. Roth. 408 U.S. 564 (1972).

Braun, Henry I. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Educational Testing Services, 2005.

Center for Educator Compensation Reform. "Dallas Principal and Teacher Incentive Pay." n.d. http://cecr.ed.gov/pdfs/profiles/Dallas.pdf.

Charles Alan Wright and Charles H. Koch, Jr. "Substantial Evidence or Reasonableness." *Federal Practice and Procedure* 33 (updated 2012): Judicial Review § 8333.

Chemerinsky, Erwin. *Constitutional Law: Principles and Policies.* Third Edition. New York: Aspen Publishers, 2006.

City of Cleburne v. Cleburne Living Center, Inc. 473 U.S. 432 (1985).

Clark v. Whiting. 607 F.2d 634 (4th Circuit, 1979).

Collins v. Faith School District No. 46-2. 574 N.W.2d889 (Supreme Court of South Dakota, 1998).

Corcoran, Sean P. *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.* Annenberg Institute for School Reform at Brown University, 2010.

D. Aaronson, et al. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (2007): 95.

D. Boyd, et al. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy* 1 (2006): 176, 177.

Day-Brite Lighting, Inc. v. Missouri. 342 U.S. 421 (1952).

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

Dillon, Sam. "Formula to Grade Teachers' Skill Gains Acceptance, and Critics." *The New York Times*, August 31, 2010: A1.

Douglas B. Downey, et al. "Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year." *American Sociological Review* 69 (2004): 613.

Economic Policy Institute. "Problems with the Use of Student Test Scores to Evaluate Teachers." Briefing Paper, 2010.

Eric A. Hanushek, Steven G. Rivkin. "Generalizations About Using Value-Added Measures of Teacher Quality." 2010.

Fisher v. Snyder. 476 F.2d 375 (8th Circuit, 1973).

Fox v. San Francisco Unified School District. 245 P.2d 603 (California Court of Appeals, 1952).

Gunther, Gerald. "Foreword: In Search of Evolving Doctrine on a Changing Court: A Model for a Newer Equal Protection." *Harvard Law Review* 86 (1972): 1.

Hamilton, Justin. "Delaware and Tennessee Win First Race to the Top Grants." United States Department of Education, 2010.

Hansen, D. Goldhaber and M. *Is it Just a Bad Class? Assessing the Stability of Measured Teacher Performance.* Working Paper No. 20085, Seattle: Center of Reinventing Public Education, 2008.

Hollingsworth v. Board of Education of School District of Alliance. 303 N.W.2d 506 (Supreme Court of Nebraska, 1981).

Houston Independent School District. "Value-Added Analysis Frequently Asked Questions." February 24, 2010. http://www.houstonisd.org/HISDConnectEnglish/Images/PDF/ValueAdded_FAQ_0209.pdf.

In re Proposed Temrination of James E. Johnson's Teaching Contract with Independent School District No. 709. 451 N.W.2d 343 (Minnesota Court of Appeals, 1990).

Jacobsen, Gene S. "The Dismissal and Non-Reemployment of Teachers." *Journal of Law and Education* 1 (1972): 435.

Jennifer L. Steele, Laura S. Hamilton, and Brian M. Stecher. *Incorporating Student Performance Measures into Teacher Evaluation Systems.* RAND, 2010.

Jonah E. Rockoff, et al. *Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools.* Working Paper No. 16240, National Bureau of Economic Research, 2011.

Karl L. Alexander, et al. "Lasting Consequences of the Summer Learning Gap." *American Sociological Review* 72 (2007): 167.

Koretz, Daniel. "A Measured Approach: Value-Added Models Are a Promising Improvement, But No One Measure Can Evaluate Teacher Performance." *American Educator*, Fall 2008: 18.

Lee v. Tuscaloosa County Board of Education. 591 F.2d 324 (5th Circuit, 1979).

Louis Fischer, David Schimmel, and Leslie R. Stellman. *Teachers and the Law.* Seventh Edition. Pearson, 2007.

McGowan v. Maryland. 366 U.S. 420 (1961).

Mellon, Ericka. "Statistical Tool for Rating Faculty Could Be Adopted Today." *Houston Chronicle*, May 11, 2011.

Miller v. Houston Independent School District. 51 S.W.3d 676 (Texas Appeals Court, Houston (1st District), 2001).

National Council on Teacher Quality. "State of the States: Trends and Early Lesson on Teacher Evaluation and Effectiveness Policies." 2011.

"No Child Left Behind Act of 2001, 20 U.S.C.A. § 9101(23)." 2001.

Ohio Department of Education. "Understanding Ohio's Accountability System 2009-2010." http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=117&ContentID=91231&cONTENT=91251 (accessed April 25, 2012).

Operation Public Education, The Center for Greater Philadelphia. *Value-Added Assessment.* The Center for Greater Philadelphia. n.d. http://www.cgp.upenn.edu/ope_value.html (accessed April 25, 2012).

—. *Value-Added Assessment in Ohio.* n.d. http://www.cgp.upenn.edu/ope_ohio.html (accessed April 25, 2012).

—. *Value-Added Assessment in Pennsylvania.* n.d. http://www.cgp.upenn.edu/ope_pa.html (accessed April 25, 2012).

Pennsylvania Department of Education. "Pennsylvania Value Added Assessment System." n.d. http://www.portal.state.pa.us/portal/server.pt/community/pa_value-added_assessment_system_(pvaas)/8751 (accessed April 25, 2012).

Perry v. Sindermann. 408 U.S. 593 (1972).

Race to the Top Applications. n.d. http://www2.ed.gov/programs/racetothetop/index.html (accessed 25 2012, April).

Raj Chetty, John N. Friedman, Jonah E. Rockoff. *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.* Working Paper No. 17699, National Bureau of Economic Research, 2011.

RAND. "The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness." 2004.

Rice, Jennifer King. *Understanding the Effectiveness of Teacher Attributes.* Economic Policy Institute, 2003.

Roger Goddard, et al. "Collective Teacher Efficacy: Its Meaning, Measure, and Impact on Student Achievement." *American Education Research Journal* 37 (2000): 479.

Romer v. Evans. 517 U.S. 620 (1996).

Scheelhaase v. Woodbury Central Community School District. 488 F.2d 237 (8th Circuit, 1973).

Scheelhaase v. Woodbury Central Comunity School District. 349 F.Supp. 988 (Northern District of Iowa, 1972).

Tennessee Deparment of Education. "Tennesssee Value-Added Assessment System - TVAAS." n.d. http://www.tn.gov/education/assessment/test_results.shtml (accessed April 25, 2012).

The New Teacher Project. "Teacher Evaluation 2.0." 2010.

The New Teacher Project. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness." 2009.

Town of Castle Rock v. Gonzales. 545 U.S. 748 (2005).

Trustees, Missoula County School District No. 1 v. Anderson. 757 P.2d 1315 (Supreme Court of Montana, 1988).

Turque, Bill. "'Creative...Motivating' and Fired." *The Washington Post*, March 6, 2012.

—. "More than 200 D.C. Teachers Fired." *The Washington Post*, July 15, 2011.

—. "Teachers in Court Fight over Evaluations." *The Washington Post*, July 1, 2011.

United States Department of Education. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." 2010.

United States Department of Education. "Measuring Teacher Effectiveness Using Growth Models: A Primer." 2011.

—. "Race to the Top Program Executive Summary." November 2009.

Va. Code Ann. § 22.1-303. Westlaw, 2012.

Wadhwani, Neelum J. "Rational Reviews, Irrational Results." *Texas Law Review* 84 (2006): 801.

Williamson v. Lee Optical. 348 U.S. 483 (1955).