FROM INSPECTION, SUPERVISION, AND OBSERVATION TO VALUE-ADDED EVALUATION: A BRIEF HISTORY OF U.S. TEACHER PERFORMANCE EVALUATIONS

Jodi Wood Jewell[*]

ABSTRACT

The debate over the appropriate model for use in K–12 teacher evaluation has increased as the language of school and teacher accountability has gained traction. The current school reform movement stresses greater teacher accountability for student learning. Constituents are no longer satisfied with traditional evaluation forms; instead, they are increasingly requesting demonstrable improvement on standardized tests. While high-stakes testing has been a feature of the educational landscape for some time, it is now being embraced not just as a measure of student learning and institutional worth, but as a tool to assess teacher quality and fitness.

As a result, teacher evaluation systems have evolved to measure teacher performance in a variety of ways, including connecting student exam performance to teacher effectiveness. The challenges implicit in adopting new teacher evaluation methods can be seen when viewing such policies from a historical perspective. This Article shows the slow development of teacher performance reviews over time; it begins with a history of teacher evaluation from the colonial era to the present, discusses the introduction of Valued Added Modeling (VAM) to ensure student achievement, provides an overview of current teacher evaluation models and concludes by identifying policy considerations in creating accurate, comprehensive and fair evaluation systems.

TABLE OF CONTENTS

* Associate Professor of Law, University of La Verne College of Law. Thanks to the editors of the *Drake Law Review* for their valuable suggestions.

I. INTRODUCTION

On July 16, 2015, the education advocacy group Students Matter filed a lawsuit against 13 California school districts in an effort to force those districts to include students' standardized test scores in teacher performance evaluations used to determine retention and promotion.[1] The suit alleged that the listed California school districts violated a state law that required objective measures—such as standardized test scores—be used to evaluate teachers.[2] The plaintiffs contended that collective bargaining agreements between teachers' unions and school districts prevented student test scores

1. *See* Verified Petition for Writ of Mandate or Other Appropriate Relief; Verified Complaint for Injunctive & Declaratory Relief at 3–6, Doe v. Antioch Unified Sch. Dist., No. 15-1127 (Super. Ct. Cal. July 16, 2015) [hereinafter Verified Petition, Doe v. Antioch]; *Doe v. Antioch*, STUDENTS MATTER, http://studentsmatter.org/case/doe-v-antioch/ (last visited Jan. 19, 2017).

2. Verified Petition, Doe v. Antioch, *supra* note 1, at 6–7; *see also* Howard Blume, *Group Sues 13 School Districts for Not Using Test Scores in Teacher Evaluations*, L.A. TIMES (July 16, 2015), http://www.latimes.com/local/lanow/la-me-ln-suit-teacher-evaluations-20150716-story.html. In an earlier case, *Doe v. Deasy*, a Los Angeles district court judge ruled that the Los Angeles Unified School District must incorporate student test scores into the teacher evaluation process but stated that the district had a great deal of discretion in determining how to do so. Judgment Granting Petition for Writ of Mandate at 1–2, 5, Doe v. Deasy, No. BS134604, 2012 WL 12529445, at *1–2 (Cal. Super. Ct. July 24, 2012); Tentative Decision on Petition for Writ of Mandate at 24–25, Doe v. Deasy (Cal. Super. Ct. June 11, 2012) (No. BS 134604), http://blogs.edweek.org/edweek/teacherbeat/Doe%20v%20%20Deasy%20Tentative%20Ruling%20Compact%20PDF.pdf.

from being used in teacher evaluations and thus, violated the state's 1971 Stull Act,[3] which requires school districts to include in their evaluation processes "[t]he progress of pupils toward the . . . academic content standards as measured by state adopted criterion referenced assessments."[4]

The plaintiffs in *Doe v. Antioch* argued that "ineffective teachers" in the classroom delayed the learning process of students assigned to such teachers; they further argued that using student standardized test scores as one measure of the teacher evaluation process would best identify teachers who are "exemplary," as well as those who are "struggling."[5] While the lawsuit did not specifically require that a certain percentage of a teacher's evaluation be derived from test scores, Joshua S. Lipshutz, an attorney for Students Matter, identified—based on external studies—that 30 to 40 percent was the ideal weight for such scores.[6]

The matter was heard on July 29, 2016, and consisted of documents filed with the court and oral argument; no testimony was heard by agreement of the parties.[7] On September 19, 2016, the trial court ruled that the Stull Act did not specifically require the incorporation of student achievement scores into individual teacher evaluations.[8] Instead the court found that the Stull Act requires California school districts to include student test data only if it "reasonably relates to the progress of pupils towards . . . the state adopted academic content standards as measured by [standardized tests]."[9] The court defined "reasonably relates" as the "discretion to determine what is reasonable in [a] complex situation" and permitted districts to use their "own judgment or opinion concerning the act's propriety or impropriety."[10] The court went on to state that "the law affords discretion to the official to determine what is reasonable under the circumstances," making mandamus

3. *See* Verified Petition, Doe v. Antioch, *supra* note 1, at 1, 5; *see also* Act of July 20, 1971, ch. 361, § 40, 1971 Cal. Stat. 720, 726–27 (codified as amended at CAL. EDUC. CODE §§ 44660–44665 (West 2017)).

4. Verified Petition, Doe v. Antioch, *supra* note 1, at 6 (first alteration in original) (emphasis omitted) (quoting CAL. EDUC. CODE § 44662(b)(1) (2015)).

5. *Id.* at 4–6.

6. Blume, *supra* note 2.

7. Doe v. Antioch Unified Sch. Dist., No. MSN15-1127, slip op. at 1 (Cal. Super. Ct. Sept. 19, 2016), http://www.cc-courts.org/general/docs/Doe_v_Antioch_Final_Opinion_and_Order.pdf.

8. *Id.* at 39–40.

9. *Id.* (alteration in original) (quoting CAL. EDUC. CODE § 44662(b)(1)).

10. *Id.* (quoting Weinstein v. Cty. of Los Angeles, 188 Cal. Rptr. 3d 557, 573 (Ct. App. 2015)).

inappropriate.[11] The court concluded by noting the "many complicated factors that bear on whether and how student test scores might reasonably relate to a teacher's performance" and left the determination of how to measure those factors to individual districts.[12]

Doe v. Antioch is just one example of the recent movement to assure higher teaching quality through greater individual teacher accountability. Increasingly, the transformation of existing teacher evaluation and tenure policies is being shaped by lawsuits filed by parents and special interest groups. *Doe v. Antioch* is not the first time Students Matter has filed suit to promote change in the educational landscape; the group filed another case to increase teacher quality in 2014.[13] In that case, *Vergara v. California*, the plaintiffs argued that the provisions of the California Education Code addressing teacher tenure and dismissal were unconstitutional,[14] as those provisions denied students equal protection by perpetuating a system that negatively impacted students unlucky enough to be placed in a classroom with a "grossly ineffective" teacher.[15] The *Vergara* plaintiffs relied on "value-added" modeling (VAM) to demonstrate that poor and minority students had a greater likelihood than other California students of being placed in a classroom with a low-performing teacher since schools with large populations of such students had correspondingly large numbers of ineffective teachers.[16] The study relied upon by *Vergara* used a data set including 2.5 million students who were tested in math and English between 1989 and 2009.[17] The test score data accounted for key family characteristics and other variables that might potentially influence student learning.[18]

11. *Id.* at 40 (comparing ministerial duties with those that require the exercise of discretion or judgment).

12. *Id.*

13. Vergara v. State, 209 Cal. Rptr. 3d 532, 538 (Ct. App. 2016); Vergara v. California, STUDENTS MATTER, http://studentsmatter.org/case/vergara/ (last visited Jan. 23, 2017).

14. *Vergara*, 209 Cal. Rptr. 3d at 539–40 (citing CAL. EDUC. CODE § 44929.21(b) (teacher tenure policy); *id.* §§ 44934, 44938(b), 44944 (dismissal and suspension policies); *id.* § 44955 (seniority policy)).

15. *Id.* at 540, *rev'g*, No. BC484642, 2014 WL 6478415 (Cal. Super. Ct. Aug. 27, 2014).

16. *Id.* at 542, 544–45; *see also* Vergara v. State, 2014 WL 6478415, at *4, *7.

17. *See* Vergara v. State, 2014 WL 6478415, at *4; Raj Chetty et al., *Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates*, 104 AM. ECON. REV. 2593, 2593–94 (2014) [hereinafter Chetty et al., *Measuring the Impacts of Teachers I*].

18. Chetty et al., *Measuring the Impacts of Teachers I*, *supra* note 17, at 2594.

Students Matter prevailed in the *Vergara* trial court case.[19] The trial court agreed that teacher tenure and dismissal provisions were unconstitutional; it found persuasive that "a single year in a classroom with a grossly ineffective teacher costs students $1.4 million in lifetime earnings per classroom."[20] Experts testifying in the case further stated that Los Angeles Unified School District (LAUSD)[21] students taught by teachers ranking "in the bottom 5 [percent] of competence lose 9.54 months of learning in a single year."[22] Judge Rolf M. Treu, who heard *Vergara* in the trial court, found the statutes unconstitutional, stating that tenure and "last in, first out"[23] policies resulted in "a real and appreciable impact on students' fundamental right to equality of education and . . . impose[d] a disproportionate burden on poor and minority students."[24]

Although the decision was overturned by the California Court of Appeal,[25] that court reviewed *Vergara* solely as a facial challenge to the constitutionality of the California Education Code as applied to teacher tenure and dismissal.[26] Thus, the court "consider[ed] only the text of the measure itself, not its application to the particular circumstances of an individual."[27]As a result, while the court found that the statutes could be applied in such a way that low-income and minority students could encounter higher numbers of "grossly ineffective" teachers during their K–12 education, it ultimately decided that the statutes were facially valid.[28] The court indicated that the task of ensuring effective teaching properly belonged with school districts, although it chastised administrators for making "deplorable staffing decisions" that had a "deleterious impact on

19. Vergara v. State, 2014 WL 6478415, at *7.

20. *Id.* at *4, *7.

21. LAUSD is the second largest public school district in the country; its 2015–2016 enrollment data shows that it serves 639,337 children in 1,005 schools and educational settings. *District Summary: Los Angeles Unified*, ED-DATA EDUC. DATA P'SHIP, http://www.ed-data.org/district/Los-Angeles/Los-Angeles-Unified (last visited Feb. 25, 2017).

22. Vergara v. State, 2014 WL 6478415, at *4.

23. CAL. EDUC. CODE § 44955(c) (West 2017) mandates that teachers "shall be terminated in the inverse of the order in which they were employed."

24. Vergara v. State, 2014 WL 6478415, at *4 (emphasis omitted).

25. Vergara v. State, 209 Cal. Rptr. 3d 532, 538 (Ct. App. 2016).

26. *Id.* at 550–51.

27. *Id.* at 550 (quoting Tobe v. City of Santa Ana, 892 P.2d 1145, 1152 (Cal. 1995)).

28. *Id.* at 557–58 (reversing the trial court because plaintiffs could not provide evidence sufficient for a *facial* challenge).

poor and minority students."[29]

Doe v. Antioch and its predecessor, *Vergara*, both demonstrate the resolve of the public, parents, and special interest groups to effect change in the way schools and teachers are being evaluated and held accountable for student achievement. Hiring and retaining high-quality teachers has become essential with students in the United States regularly performing below their foreign counterparts on international exams.[30] The gap between U.S. students and students in other nations also widens as students progress to higher grades—by age 15, U.S. students rank "38th out of 71 countries in math and 24th in science."[31] Because national economic growth is tied to an educated workforce and leads to increased gross domestic product over time,[32] a stronger educational model is required. Both *Vergara*[33] and *Doe v. Antioch*[34] illustrate public dissatisfaction with current educational outcomes: K–12 students are not achieving key learning objectives and thus are unprepared to compete in a global economy. Both cases also underscore the public perception that education is unresponsive to the needs of the children whose lives it shapes.

While society has perceived U.S. education as being inadequate almost since public education began to be provided by the states, the debate over what model is appropriate for evaluating K–12 teaching has increased in recent years as the language of school and teacher accountability has gained traction. While the final decision on what constitutes appropriate teacher

29. *Id.* at 557.

30. *See* EDWARD H. HAERTEL, RELIABILITY AND VALIDITY OF INFERENCES ABOUT TEACHERS BASED ON STUDENT TEST SCORES 4 (2013), https://www.ets.org/Media/Research/pdf/PICANG14.pdf.

31. Drew DeSilver, *U.S. Students' Academic Achievement Still Lags That of Their Peers in Many Other Countries*, PEW RES. CTR. (Feb. 15, 2017), http://www.pewresearch.org/fact-tank/2015/02/02/u-s-students-improving-slowly-in-math-and-science-but-still-lagging-internationally/; s*ee also* ANINDITA SEN ET AL., NAT'L CTR. FOR EDUC. STATISTICS, U.S. DEP'T OF EDUC., COMPARATIVE INDICATORS OF EDUCATION IN THE UNITED STATES AND OTHER G8 COUNTRIES: 2004, at 28, 48 (2005), https://nces.ed.gov/pubs2005/2005021.pdf (finding U.S. fourth graders ranked second highest in literacy, but U.S. 15-year-olds displayed a lower engagement in reading than their peers in other comparable countries).

32. Eric A. Hanushek, *Teacher Deselection*, *in* CREATING A NEW TEACHING PROFESSION 169 fig.8.3 (Dan Goldhaber & Jane Hannaway eds., 2009).

33. *See Vergara*, 209 Cal. Rptr. 3d at 557; Vergara v. State, No. BC484642, 2014 WL 6478415, at *2 (Cal. Super. Ct. Aug. 27, 2014), *rev'd*, 209 Cal. Rptr. 3d 532 (2016).

34. *See* Verified Petition, Doe v. Antioch, *supra* note 1, at 4–5.

evaluation rests with the states,[35] pressure from the U.S. Department of Education and other interested parties is driving change to both teacher tenure policies and evaluation.[36]

The dissatisfaction with the state of public education has led to a reform movement that stresses greater teacher accountability for student learning.[37] Constituents are no longer satisfied with traditional evaluation forms—instead they are increasingly demanding demonstrable improvement on standardized tests.[38] While high-stakes testing has been a feature of the educational landscape for some time, it is now being embraced, not just as a measure of student learning and institutional worth, but as a tool to calculate teacher quality and fitness.[39] As a result, schools across the country have embraced, or are moving toward, teacher evaluation models that measure teacher performance in a wide variety of ways; unsurprisingly, a number of these models propose connecting student standardized test performance to teacher job performance.[40]

The challenges implicit in adopting new teacher evaluation policies can be seen when viewing such policies from a historical perspective. This perspective shows the slow development of evaluation over time and contrasts it with the relatively recent calls to associate student test scores with teacher evaluation scores. This Article creates a historical framework for discussion; it begins with a history of teacher evaluation from the colonial era to the present,[41] discusses the relatively recent introduction of "value-added" models (VAMs) as mechanisms to ensure student achievement,[42] provides an overview of current state teacher evaluation models,[43] and

35. U.S. DEP'T OF EDUC., RACE TO THE TOP PROGRAM: EXECUTIVE SUMMARY 2–4 (2009) [hereinafter RACE TO THE TOP], https://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

36. *See, e.g.*, *id.* at 9.

37. *See* HAERTEL, *supra* note 30.

38. EVA L. BAKER ET AL., ECON. POLICY INST., BRIEFING PAPER NO. 278, PROBLEMS WITH THE USE OF STUDENT TEST SCORES TO EVALUATE TEACHERS 1 (2010), http://www.epi.org/files/page/-/pdf/bp278.pdf.

39. *See, e.g.*, *id.*; HAERTEL, *supra* note 30.

40. *See, e.g.*, Verified Amended Complaint at 9–16, Davids v. State, No. 101105/14 (N.Y. Sup. Ct. July 25, 2014) (arguing that New York teacher tenure statutes deprived students of their state constitutional right to sound education because of administrative failure to find, assess, and dismiss incompetent teachers).

41. *See infra* Parts II–III.

42. *See infra* Part IV.

43. *See infra* Part V.

identifies potential barriers to creating an accurate, comprehensive, and fair evaluation system.[44]

II. A HISTORY OF EDUCATION AND TEACHER EVALUATION FROM THE COLONIAL ERA TO NO CHILD LEFT BEHIND

The national discussion surrounding individual teacher accountability—and its counterpart, school accountability—is relatively new. Prior to 1965, decisions about curriculum content, teaching methodologies, and teacher evaluations were primarily left to the states.[45] Since there was no federal right to education, students had to rely on the educational rights provided by the state constitution of the state in which they resided. This led to uneven effects as educational opportunities and quality of instruction varied on multiple levels; there was great variety between states, between districts within a state, and between local schools. When the Elementary and Secondary Education Act (ESEA) was enacted in 1965, its goal was to erase these differences and to provide greater educational parity to students living in poor districts.[46] The ESEA accelerated changes in education by expanding the federal government's reach into K–12 education as states exchanged their autonomy for federal funding.[47]

Early U.S. education was a purely private model; the children of the wealthy were tutored at home, and their sons were later sent to universities.[48] The earliest change to this model occurred when public schools were established in 1647 in Massachusetts in an effort to equip children with the "ability to read and understand the principles of religion and capital laws of this country."[49] Towns of more than 50 homes were required to hire someone to teach reading and composition to children; larger towns of 100 or more

44. *See infra* Part VI.

45. WAYNE J. URBAN & JENNINGS L. WAGONER, JR., AMERICAN EDUCATION: A HISTORY 374 (Routledge, 4th ed. 2009) (1996).

46. *See* Elementary and Secondary Education Act of 1965 (ESEA), Pub. L. No. 89-10, sec. 2, § 201, 79 Stat. 27, 27.

47. *See id.* sec. 2, §§ 202, 203, 79 Stat. at 27–30.

48. ALAN R. SADOVNIK ET AL., EXPLORING EDUCATION: AN INTRODUCTION TO THE FOUNDATIONS OF EDUCATION 66 (2d ed. 2001) (1994).

49. *Id.* at 67 (quoting 2 RECORDS OF THE GOVERNOR AND COMPANY OF THE MASSACHUSETTS BAY IN NEW ENGLAND 6 (Nathaniel B. Shurtleff ed., William White 1853)).

were required to establish secondary schools.[50] Massachusetts was a pioneer in early elementary and secondary education; outside of New England, schooling opportunities varied widely.[51]

The student population in these early institutions was uniform rather than diverse, as the majority of students had European ancestry.[52] Very few schools accepted African Americans; slave owners restricted educational access, believing it was unnecessary and would lead to insurrection.[53] The exception to this lack of educational access for African Americans was a limited number of schools that were run by Quakers and Anglicans.[54] There was also limited educational opportunity in this era for Native Americans.[55] While some religious groups attempted to provide education, initial efforts were unsuccessful, and Native Americans were treated as peripheral to the educational system.[56]

Early U.S. school structure generally consisted of a teacher instructing a single classroom of students ranging from the primary grades to the upper grades.[57] There was strict discipline and a strong focus on memorization and recitation.[58] Schools were administered by local community leaders who determined the academic goals for their schools and decided whether those had been met; the family, society, and religious institutions were of primary significance.[59] The curriculum components emphasized varied depending upon the community in which the school was located; schools mirrored the values, religions, and ethnicities of the colonists in each region.[60]

50. *Id.*

51. URBAN & WAGONER, *supra* note 45, at 44–45, 57.

52. *See* SADOVNIK ET AL., *supra* note 48; URBAN & WAGONER, *supra* note 45, at 54–55.

53. SADOVNIK ET AL., *supra* note 48, at 70.

54. *Id.*

55. *See id.*

56. *Id.*

57. The earliest U.S. public school was the Boston Latin School established in 1635 in Boston, Massachusetts. *BLS History*, BOS. LATIN SCH., www.bls.org/apps/pages/index.jsp?uREC_ID=206116&type=d (last visited Feb. 2, 2017). In some instances, students were taught in the homes of widows or older women. Such schools were called dame schools. SADOVNIK ET AL., *supra* note 48, at 68.

58. SADOVNIK ET AL., *supra* note 48, at 69.

59. Saundra J. Tracy, *How Historical Concepts of Supervision Relate to Supervisory Practices Today*, 68 CLEARING HOUSE 320, 320 (1995); *see* SADOVNIK ET AL., *supra* note 48, at 67–69.

60. SADOVNIK ET AL., *supra* note 48, at 68–69.

Consequently, many educational goals were tied to religion.[61] These early schools were open to the public, and costs varied.[62] In some cases, parents sending children to school paid a fee; in others, the cost was borne by the wider community.[63] Regardless, there was less focus on teacher quality and greater focus on moral fitness and curriculum suitability.

The teacher evaluation process in early colonies was primarily a system of inspection. Many teachers in these early years of public education did not have a great deal more education than their students.[64] A teacher's effectiveness was evaluated through community and religious mores[65] rather than through quality of instruction and student achievement. Community leaders visited schools to make sure their preferred curriculum was being addressed, reviewed whether the teacher was maintaining discipline in the classroom, and made sure that the teacher was providing appropriate physical maintenance of the premises. [66] Control over the school and the teacher was concentrated in local community authorities; there was little state, and no federal, oversight.[67] The remedy for a teacher's failure to meet expectations was typically dismissal rather than training and feedback to help the teacher improve.[68] Teachers were expected to be immediately responsive to the direction of community leaders who had the power to terminate the teacher for any infraction.[69]

Education remained under state and local control after the establishment of the federal government; however, the educational landscape began to change in the 1800s. Between 1820 and 1860, the Industrial Revolution increased urban development in the United States and attracted immigrant workers.[70] Since access to education was limited,

61. *Id.* at 69.

62. *See* URBAN & WAGONER, *supra* note 45, at 45 (discussing early educational laws).

63. *See id.*

64. In fact, "two-thirds of the schoolmasters in Maryland were either indentured servants or convicts." SADOVNIK ET AL., *supra* note 48, at 69 (quoting LOUIS B. WRIGHT, THE CULTURAL LIFE OF THE AMERICAN COLONIES 1607–1763, at 101 (1957)).

65. Tracy, *supra* note 59.

66. *Id.*

67. *See id.* (noting "the strong American belief in local, lay control of education").

68. *See* ROBERT J. MARZANO ET AL., EFFECTIVE SUPERVISION: SUPPORTING THE ART AND SCIENCE OF TEACHING 12 (2011); Tracy, *supra* note 59, at 321.

69. Tracy, *supra* note 59, at 321.

70. SADOVNIK ET AL., *supra* note 48, at 70–71.

teaching and instructional materials were varied, and the school year was relatively short, the majority of Americans were illiterate.[71] Teachers also lacked qualifications, and textbook recitation remained the primary method of learning.[72] However, as the numbers of schools increased and education became more readily obtainable, the curriculum began to shift from community-approved subjects and religious training to more academic subject matter.[73] This resulted in the need for better-educated teachers with better training by an expert, or "principal," teacher.[74]

Because of this population shift, the structure and delivery of education began to separate into two distinct organizational layers. The top layer was comprised of school administrators and university professors who were well-paid and who broke ground by setting policy and planning the curriculum and course of public education.[75] The lowest layer consisted of primarily female teachers who taught the classes but who were paid less and who had little opportunity for input into charting the course of education.[76] After the 1840s, the call for education "reform" primarily took the form of decreasing teacher autonomy and increasing administrative control.[77]

The reliance on community leaders to inspect and evaluate teachers was diminished after the 1800s as schools moved to an administrative model that continued to rely on inspection but recognized the importance of teacher training.[78] The growing need for expert teachers and administrators who were familiar with more academic subject matter and capable of training others was the precursor of later administrative educational models.[79] To meet these needs, school districts were organized, and experienced teachers were hired to train subordinates; as schools became larger and duties became more complex, additional administrative roles

71. *Id.* at 71.

72. DIANE RAVITCH, LEFT BACK: A CENTURY OF FAILED SCHOOL REFORMS 21 (2000) [hereinafter RAVITCH, LEFT BACK].

73. Tracy, *supra* note 59, at 320–23.

74. MARZANO ET AL., *supra* note 68, at 13.

75. Thomas S. Popkewitz, *Professionalization in Teaching and Teacher Education: Some Notes on its History, Ideology, and Potential*, 10 TEACHING & TCHR. EDUC. 1, 3–4 (1994).

76. *Id.* at 4.

77. *Id.*

78. Tracy, *supra* note 59, at 321–23.

79. *Id.*

were required.[80] Thus, superintendents and principals became responsible for observing the quality of teaching and helping teachers improve both their skill level and knowledge of varied academic subject matter.[81] While observation was introduced as an administrative practice in this era, the focus was less on dismissal and more on providing tools to help teachers increase their expertise.[82]

As the numbers of schools grew, so too did the interest in improving teacher pedagogy.[83] In 1837, Horace Mann lobbied the Massachusetts legislature to create a state board of education and to mandate school attendance.[84] His efforts began a unified drive toward educational access that eventually spread to other states.[85] Mann believed that education could transform society and provide access to opportunity.[86] However, even though Mann was progressive in his policies regarding access, curricular offerings remained limited.[87] Mann was more concerned that students learn social and moral values than that they learn any particular subject.[88]

As part of the effort to establish and improve elementary education, Mann instituted "normal schools" to provide teacher training.[89] In these schools, students honed their teaching skills through observation and feedback.[90] The first program evaluator, Cyrus Pierce, stated, "I comment upon what I have seen and heard . . . , telling them what I deem good, and what faulty, either in their doctrine or their practice, their theory or their manner."[91] This new educational model combined an administrative oversight model with observation and feedback to lay the foundation for how teachers would be trained and evaluated in the future.[92]

80. *Id.* at 321.

81. *Id.* at 321–23.

82. *See* MARZANO ET AL., *supra* note 68, at 13.

83. *See* DANA GOLDSTEIN, THE TEACHER WARS: A HISTORY OF AMERICA'S MOST EMBATTLED PROFESSION 24–25 (2014).

84. *Id.* at 23–24.

85. *See id.* at 25–30.

86. *See* SADOVNIK ET AL., *supra* note 48, at 71.

87. GOLDSTEIN, *supra* note 83, at 27–28.

88. *Id.*

89. *Id.* at 25. Interestingly, these normal schools were open only to women, as they were cheaper to employ. *Id.*

90. *Id.*

91. *Id.*

92. *Id.* at 25–26. Eventually these normal schools were brought under the aegis of state universities and transformed into colleges of education. *Id.* at 26.

As a result of Mann's efforts and the work of other educational reformers, by 1860, despite numerous challenges to the tax burden posed by public schools and concerns by Roman Catholics that schools espoused a Protestant viewpoint, elementary public schools had become an accepted part of the U.S. landscape.[93] The belief in access and education as the key to social movement and opportunity was gaining ground.

Despite this new emphasis on access to education, such opportunities for African Americans in the South continued to be limited; in fact, teaching slaves to read and write was prohibited.[94] Outside the South, educational opportunities for African Americans were inferior; they were separated into their own institutions, a practice affirmed by *Roberts v. City of Boston*, which upheld the separation of schools by race.[95]

Because of immigration patterns and the proliferation of schools post-Mann, "[b]y 1890, 95 percent of children between the ages of five and thirteen were enrolled in school," although "[l]ess than 5 percent of adolescents went to high school, and even fewer entered college."[96] School control remained concentrated at the local level; while state and federal education agencies existed, they had very little power over what occurred in schools.[97] Additionally, while administrative oversight was the accepted model for teacher evaluation, this oversight was limited.[98] For example, in 1890, the city of Baltimore had two superintendents to oversee the entire district, which employed 1,200 teachers.[99]

As the twentieth century approached, immigration patterns shifted, the gap between rich and poor became pronounced, and schools found it necessary to shift pedagogy and to plan curricular reform.[100] The influx of immigrants into urban areas brought increased student enrollment and additional challenges; for example, in 1909, 57.8 percent of enrolled students in large cities were immigrants.[101] Given this flood of students, educators

93. SADOVNIK ET AL., *supra* note 48, at 71–72. This led the Catholic Church to establish its own schools, supported by the religious community. *Id.* at 72.

94. *Id.* at 73.

95. Roberts v. City of Boston, 59 Mass. (5 Cush.) 198, 209–10 (1849); *see also* SADOVNIK ET AL., *supra* note 48, at 73.

96. RAVITCH, LEFT BACK, *supra* note 72, at 20 (footnote omitted).

97. *Id.*

98. *Id.*

99. *Id.*

100. *See* SADOVNIK ET AL., *supra* note 48, at 74.

101. *Id.*

began debating the definitive goal of education: Was the goal to provide child-centered reform and an individual educational process, or was it social efficiency, including preparing students for work? The child-centered system advocated for an individualized approach where students would be educated in a meaningful, personalized way.[102] The social efficiency model emphasized the need for "social engineering" to mold children into model citizens.[103] Finally, social efficiency stressed preparation for life as part of the school process.[104] Schools were also called on to socialize and "Americanize the diverse groups who had become citizens."[105] These challenges and alternative theories of reform, combined with other pressures, changed the face of U.S. education and forced schools away from prior practices and toward new ones.[106]

As urban areas grew, the larger numbers of schools created a bureaucracy with multiple administrative levels and little teacher autonomy.[107] The need to standardize the curriculum and create "American" citizens abridged teacher influence over academic content and "increased the monitoring and control of their work by others."[108] Along with other increased administrative tasks, more formal teacher evaluations were introduced.[109] These new administrative, higher-paying roles of superintendent and principal were largely filled by men, while the lower-paying teaching roles continued to be filled by women.[110]

Although Americans supported elementary education in this time, at the end of the nineteenth century, the support for secondary education continued to lag behind. Unlike the accessible system of public high schools Americans have today, most secondary students attended "small private academies" that "offered not only the classical curriculum of Latin, Greek, and mathematics, but also . . . history, science, and English."[111] However, the debate about the goals of education, combined with the shifting economy and larger populations in cities, drove the demand for more public high

102. *Id.* at 75–76.
103. *See id.* at 76.
104. *Id.*
105. Popkewitz, *supra* note 75, at 5–6.
106. *See* SADOVNIK ET AL., *supra* note 48, at 76–78.
107. Popkewitz, *supra* note 75, at 4; Tracy, *supra* note 59, at 321.
108. Popkewitz, *supra* note 75, at 8.
109. *Id.* at 4.
110. *Id.*
111. RAVITCH, LEFT BACK, *supra* note 72, at 25.

schools.[112] Prior to 1875, high school attendance was not required, and "fewer than 25,000 students were enrolled in public high schools. . . . [B]etween 1880 and 1920 [when all states had adopted compulsory high school attendance], 2,382,542 students attended public high schools . . . and by 1940, about 6.5 million students attended public high school."[113] The increase in high school attendance was partially due to the compulsory school age being raised during the Great Depression to limit the number of people in the job market, but high schools gradually gained mainstream acceptance as they began offering courses preparing students for future employment.[114]

Student population growth, combined with the increase in the number of schools over a relatively short time period, necessitated changes to curriculum, teaching methods, and teacher evaluation. Educators questioned whether preparing students for college through teaching the classics was the best approach, or whether schools should provide more employment training.[115] In seeking to resolve these two viewpoints, one consideration was the administrative cost of offering various curricular tracks over providing a standardized curriculum.[116] The question of uniformity versus variety was a pressing issue requiring a solution. Many schools had been providing what was essentially a liberal arts education, with math, history, science, and languages, both classical and modern, as core courses.[117] Since very few students in this time period actually attended college, some reformers argued that students would be better served by "address[ing] . . . practical concerns of daily living,"[118] while others argued such divisions in curriculum amounted to determining what a student would learn based on social class.[119] This debate about educational goals and the appropriate curriculum was common; it was frequently advanced both before this era and afterwards.

112. *Id.*

113. SADOVNIK ET AL., *supra* note 48, at 76.

114. DIANE RAVITCH, THE TROUBLED CRUSADE: AMERICAN EDUCATION, 1945–1980, at 10–11 (1983) [hereinafter RAVITCH, TROUBLED CRUSADE].

115. DIANE RAVITCH, THE SCHOOLS WE DESERVE: REFLECTIONS ON THE EDUCATIONAL CRISES OF OUR TIMES 137 (1985) [hereinafter RAVITCH, SCHOOLS WE DESERVE].

116. *See id.* at 136–37.

117. *Id.* at 138.

118. SADOVNIK ET AL., *supra* note 48, at 78.

119. RAVITCH, SCHOOLS WE DESERVE, *supra* note 115, at 139–40.

Underlying the arguments regarding curriculum was the overt belief that children of immigrants had lower intelligence and therefore were best suited for manual labor.[120] While some critics claimed that schools with vocational curricula reinforced the idea that not all students were capable of rigorous academic learning and that only the children of elite members of society should receive a classical education while poor and minority children were less capable and should be funneled into nonacademic programs,[121] still others argued against enrolling students in courses for which they "lacked the intellect" and suggested that allowing such students to attend college would undermine the social order.[122]

In combination with these new educational theories of reform and the explosion of the student population, new educational management methods emerged.[123] Between 1900 and 1920, it was proposed that teaching could be measured and made more efficient using successful business productivity methods.[124] This concept shifted teacher evaluation away from an inspection model toward increased teacher observation and the development of objective criteria to measure performance.[125]

Even though business productivity models influenced the emerging teacher evaluation model, supervisors and principals remained the tools of carrying evaluations out; their ability to assess performance accurately was presumed.[126] In contrast to the early colonial model in which teachers were expected to perform well or suffer the consequences, the objective evaluation model required teachers and administrators to work together to improve the overall quality of the teachers' skills; the goal was retention and improvement rather than dismissal.[127]

Science-based models were also seen as a source of information to draw on for curriculum design; they were seen as a more systematic and reliable way to measure learning and human progress.[128] One psychologist,

120. RAVITCH, LEFT BACK, *supra* note 72, at 56.
121. SADOVNIK ET AL., *supra* note 48, at 79.
122. RAVITCH, LEFT BACK, *supra* note 72, at 56.
123. Tracy, *supra* note 59, at 324–25.
124. *Id.* at 323.
125. *Id.*
126. *Id.*
127. PETER J. BURKE & ROBERT D. KREY, SUPERVISION: A GUIDE TO INSTRUCTIONAL LEADERSHIP 10 (2d ed. 2005).
128. Popkewitz, *supra* note 75, at 5.

Edward Thorndike, believed that responsibility for curriculum design should be taken away from educators and placed in the hands of psychologists; he hypothesized that measuring mental growth exceeded the ability of teachers.[129] The conviction that science could be used to measure learning and growth influenced the ongoing development of teacher pedagogy; it also influenced teacher training and supervision moving forward.[130]

As education moved into the post-World War II era, the question of whether federal aid should be granted to state education was raised, as it had been at multiple points in U.S. history.[131] Prior to World War II, the obstacles to federal funding were religion, race, and federal preemption of a state function.[132] However, the G.I. bill was sending greater numbers of students than ever to college.[133] Where once higher education was reserved for the upper classes, in the post-war world there was great demand for math and science skills to meet the industry and governmental needs.[134] This led to a debate about curriculum that highlighted the persistent tension between viewing schools as places that provide academic learning versus viewing schools as places that prepare students for work.[135] This caused some critics to claim that "intellectual goals" were being subordinated to "social ones."[136]

While access was not a new objective in education, after World War II, the need to address school segregation and its effect on equal educational opportunities for African-American children became imperative. When *Brown v. Board of Education* was decided in 1954, it reversed the "separate but equal" view of education and revealed the differing qualities of education offered to students.[137] The lack of school equality was not isolated in the South; instead, there was inequality in education throughout the

129. *Id.*

130. *Id.*

131. RAVITCH, TROUBLED CRUSADE, *supra* note 114, at 5–6. Some such debates occurred: in World War I regarding the need to alleviate illiteracy in military draftees; during the Great Depression, when school districts could not afford to keep schools open; and at the conclusion of World War II, when teachers from the South lobbied for federal aid to provide equal opportunity to poor and minority children. *Id.* at 5.

132. *Id.* at 5.

133. *Id.* at 14–15.

134. *Id.*

135. *Id.* at 17–18.

136. SADOVNIK ET AL., *supra* note 48, at 79.

137. Brown v. Bd. of Educ. of Topeka, 347 U.S. 483, 495 (1954), *supplemented by* 349 U.S. 294 (1955).

United States.[138] For example, in New England, the Boston School Committee had a policy of segregation that prevented African-American students from attending high-caliber schools.[139] Even when schools were not segregated by law, they frequently were segregated in fact based on neighborhood, thus limiting opportunity.[140]

As the U.S. population continued to grow, the number of school districts increased, as did the number of subjects taught: fine arts, physical education, foreign languages, and home economics were just a few additions.[141] The focus on scientific measurement of teacher performance also diminished slightly, and a more cooperative emphasis was introduced between 1930 and 1959.[142] The cooperation model of teacher evaluation had similar origins to its predecessors; it was rooted in workplace efficiency and based on the ideas proposed by the Hawthorne studies that production improved when workers were observed.[143]

When applied to teaching, the Hawthorne model suggested that if teachers were treated as valued partners in the educational process, improved teaching quality would automatically result.[144] This cooperative model was based on observation using objective criteria but provided greater teacher participation and autonomy.[145] This new model required principals and supervisors to work with teachers to achieve results and emphasized assisting, rather than directing, the teacher.[146] Positive work partnerships were encouraged as part of the teacher's professional growth.[147]

138. SADOVNIK ET AL., *supra* note 48, at 83.

139. *Id.* at 83–84.

140. *Id.* at 83 (discussing de jure and de facto segregation).

141. BURKE & KREY, *supra* note 127, at 11.

142. *Id.* at 10–11.

143. In the Hawthorne studies, factory workers were studied to determine whether changes to their working conditions increased or decreased output. C.W.M. Hart, *The Hawthorne Experiments*, 9 CANADIAN J. ECON. & POL. SCI. 150, 153–54 (1943). The surprising finding was that the intense focus on the workers for the purposes of study increased output regardless of the type of change made to the environment. *Id.* at 153–55. In addition, when workers were interviewed in a nonjudgmental way to seek their input, productivity increased as well. *Id.* at 156–59. The conclusion reached at the time was that efficiency increased when workers were consulted about their work environment and hierarchy was suspended. *Id.* at 158–59.

144. Tracy, *supra* note 59, at 323.

145. *Id.*

146. *Id.*

147. *Id.*

Unfortunately, this model was not completely successful; it occasionally resulted in less stringent (or little) classroom observation in an effort to preserve the rapport between teacher and supervisor, since the main focus was on emphasizing a teacher's value in order to increase his or her investment in the educational process.[148] Improved teaching quality was considered to be a natural consequence of this model.[149]

In the late 1950s, the federal government began to play a greater role in public education. Separate curricular tracks were common in schools; IQ, grades, and perceived ability determined whether students qualified for academic versus vocational tracks.[150] The continuing argument about whether to provide these separate tracks was moderated somewhat when the Soviet Union won the space race with the Sputnik launch in 1957.[151] This seemed to prove that an "academic" curriculum was needed to compete with other nations.[152] The Sputnik launch caused concern about the United States' lack of perceived scientific capability and opened the door to federal funding of higher education.[153] Math, science, and foreign languages were targeted by the National Defense Education Act (NDEA) as specific areas needing improvement.[154] While the NDEA only provided college access and programs, the passing of this act showed the concern the federal government had with the United States' ability to compete for scientific achievement against rival nations on the world stage.[155] The NDEA first opened a door to federal involvement in higher education and then set the stage for eventual federal involvement in K–12 education.

The Hawthorne-inspired teacher evaluation model continued in use until the 1960s, when evaluations again shifted back to an objective model that emphasized the use of quantitative research to determine the best way

148. *Id.* at 323–24.

149. *See id.*

150. RAVITCH, LEFT BACK, *supra* note 72, at 368.

151. *Id.* at 383.

152. SADOVNIK ET AL., *supra* note 48, at 79.

153. *Sputnik Spurs Passage of National Defense Education Act*, U.S. SENATE (Oct. 4, 1957), https://www.senate.gov/artandhistory/history/minute/Sputnik_Spurs_Passage_of_National_Defense_Education_Act.htm [hereinafter *Sputnik*].

154. RAVITCH, LEFT BACK, *supra* note 72, at 362. In this early federal educational program, the money was intended to benefit only gifted students, rather than targeting the wider student population. DIANE RAVITCH, NATIONAL STANDARDS IN AMERICAN EDUCATION: A CITIZEN'S GUIDE 48 (1995) [hereinafter RAVITCH, NATIONAL STANDARDS].

155. *Sputnik*, *supra* note 153.

to improve instruction.[156] Cooperation as developed by the prior model was not completely swallowed up by objective external measurements; positive teacher interaction with students and parents became a component of teacher success.[157] However, the types of objective measurements proposed became more technical, and the collection of multiple data points was stressed as a way to improve overall educational quality.[158] As part of the push to collect data, standardized test scores began to be viewed as one way to amass data and thus objectively assess instructional quality.[159]

Further changing the landscape of education was passage of the 1965 Elementary and Secondary Education Assistance Act (ESEA), which was designed to aid minority and low-income children.[160] The ESEA, with its strong connection to the Civil Rights movement, was an attempt to respond to the need to provide instruction for disadvantaged children who were being denied access to a quality education.[161] Schools became caught up in the social changes sweeping the nation; earlier educational practices and allocations of resources were viewed as failures to address prevailing societal problems.[162] Education was blamed for everything, from perpetuating cycles of poverty to the lack of qualified science and math graduates.[163]

The ESEA enlarged the once-limited federal role in education; one of the strings attached to the federal aid provided under the act was the requirement that schools receiving Title I funds agree to participate in standardized tests.[164] The ESEA was also responsible for sending $1 billion in funds directly to schools; however, misuses of the funds required Congress to amend the ESEA four times between 1965 and 1980 in order to target the needs of disadvantaged students more directly.[165]

156. BURKE & KREY, *supra* note 127, at 12.

157. *See* Tracy, *supra* note 59, at 324 (noting human relations continued to be a focus, despite increased reliance on objective external measurements).

158. BURKE & KREY, *supra* note 127, at 12.

159. Tracy, *supra* note 59, at 324.

160. ESEA, Pub. L. No. 89-10, sec. 2, § 2, 79 Stat. 27, 27.

161. SADOVNIK ET AL., *supra* note 48, 79–80.

162. *Id.*

163. Larry Cuban, *The Open Classroom*, EDUC. NEXT, Spring 2004, at 69, 69. One interesting side effect of the upheaval was "open education," proposed as the latest tool for educational success. *Id.* Open education had no lesson plans or testing. *Id.* at 70. Instead, students worked at their own pace at "interest centers" where teachers directed their learning. *Id.* There were no desks or walls, and students were free to roam. *Id.*

164. RAVITCH, NATIONAL STANDARDS, *supra* note 154, at 47.

165. Janet Y. Thomas & Kevin P. Brady, *The Elementary and Secondary Education*

In the 1970s, evaluation systems continued to evolve into a model known as "clinical supervision."[166] Clinical supervision incorporated a multi-phase process that required the supervisor and the teacher to plan, observe, analyze, and discuss the teacher's "professional practice."[167] This model required objective measurements be combined with pre-observation, observation, and post-observation meetings where teachers and administrators worked together to improve overall teaching quality and classroom management.[168] The burden on administrators in this model was greater than in prior models, as administrators had to be experienced data gatherers and analysts with strong management skills.[169] The practice gained popularity and by 1980 was the system of choice in most schools.[170]

In 1979, against the backdrop of change, the federal government took another step toward engaging further in the K–12 educational process by creating the U.S. Department of Education.[171] While conservative legislators were opposed to federal interference in what they saw as a state and local issue, those objections were overcome by narrowing the focus to creating educational equality for all students; local control over educational standards (including teacher evaluations) was retained.[172] The attempt to increase federal involvement in K–12 education lessened slightly in 1980 when President Ronald Reagan, as part of a more conservative agenda, reduced federal involvement in public education by limiting funding and decreasing regulation.[173] However, this did not eliminate federal concern with elementary and secondary education; it was increasingly discussed on the national stage as policymakers became more concerned about consistent academic quality year after year.[174]

These concerns came to the forefront with the publication of *A Nation*

Act at 40: Equity, Accountability, and the Evolving Federal Role in Public Education, 29 REV. RES. EDUC. 51, 52–53 (2005).

166. MARZANO ET AL., *supra* note 68, at 17–18.

167. *Id.* at 18.

168. *See id.*

169. *See* Tracy, supra note 59, at 324.

170. MARZANO ET AL., *supra* note 68, at 17.

171. D.T. STALLINGS, A BRIEF HISTORY OF THE UNITED STATES DEPARTMENT OF EDUCATION: 1979–2002, at 4 (2002), https://childandfamilypolicy.duke.edu/pdfs/pubpres/BriefHistoryofUS_DOE.pdf.

172. *Id.*

173. Thomas & Brady, *supra* note 165, at 53.

174. *Id.*

at Risk: The Imperative for Educational Reform in 1983.[175]*A Nation at Risk* warned that U.S. education was not preparing students for the era in which they lived.[176] The report emphasized the need for "higher academic standards, increased student course requirements, a longer school day, and significant changes in the training and retention of teachers."[177] Rather than separating students into different curricular tracks, the report recommended that all students should have access to quality education.[178] Too much choice, according to the report, was resulting in students taking fewer rigorous academic courses.[179] The report convinced states both that educational reform was necessary and that the federal government would inevitably play a role in the solution. [180]

To improve teaching—one of the four main areas flagged for improvement—*A Nation at Risk* encouraged the professionalization of teachers.[181] This need was further recognized in the Rand Corporation's study of 32 school districts across the United States to isolate effective teacher evaluation practices.[182] The study concluded that evaluation processes must meet the needs of students located in the geographic area and align with the stated educational goals of districts.[183] It further determined that districts needed to provide adequate time and allocate enough resources to evaluation to make it successful.[184] Additionally, the report stated that: evaluation goals needed to be predetermined and used to create the evaluation system, rather than the other way around;[185] evaluation systems should target the problems of specific districts or schools to justify

175. Nat'l Comm'n on Excellence in Educ., *A Nation at Risk: The Imperative for Educational Reform*, 84 ELEMENTARY SCH. J. 112 (1983) [hereinafter *A Nation at Risk*].

176. *Id.* at 112–14.

177. Thomas & Brady, *supra* note 165, at 53–54.

178. *See A Nation at Risk*, *supra* note 175, at 123–24.

179. *Id.* at 122–23 ("In effect, we have a cafeteria-style curriculum in which the appetizers and desserts can easily be mistaken for the main courses.").

180. Michael Heise, *Goals 2000: Educate America Act: The Federalization and Legalization of Educational Policy*, 63 FORDHAM L. REV. 345, 346 (1994).

181. *See A Nation at Risk*, *supra* note 175, at 126–27. Such professionalization would increase status and provide greater autonomy and responsibility for student learning growth. Popkewitz, *supra* note 75, at 3.

182. ARTHUR E. WISE ET AL., TEACHER EVALUATION: A STUDY OF EFFECTIVE PRACTICES 4 (1984), http://www.rand.org/content/dam/rand/pubs/reports/2006/R3139.pdf.

183. *Id.* at 66.

184. *Id.* at 68.

185. *Id.* at 70.

the investment of time and money;[186] and teacher participation was required both in developing evaluation standards and in providing feedback to colleagues.[187] Teachers surveyed as part of the study identified additional areas as needing improvement: the use of assorted evaluation systems, the lack of principals with the "resolve and competence" to review teachers accurately, the nonexistence of appropriate evaluator training, and the need to close the loop between teachers receiving but not incorporating feedback.[188]

A Nation at Risk and the Rand study began an important national conversation about improving public education, refining curricular offerings, and motivating effective teaching practices. Pressure mounted on states to improve academic quality and to use standardized testing to measure student progress.[189] The reauthorization of the ESEA in 1988 attempted to address these needs.[190] It emphasized the importance of allocating resources to underserved student populations and required states to use standardized testing and reporting to ascertain whether students were meeting educational goals.[191] While it took time for these ideas to gain traction, federal policymakers continued to press for national educational standards measured by objective testing.[192]

As teacher evaluation moved into the 1980s, the adoption and use of combined objective and cooperative models continued. One of the key influences on teacher evaluation models during this time was Madeline Hunter's seven-step model of mastery learning, which evolved into a teacher evaluation structure in many states.[193] Hunter's model prescribed observation and script-taping as essential components of teacher growth; it

186. *Id.* at 73–74.

187. *Id.* at 76–77.

188. MARZANO ET AL., *supra* note 68, at 23.

189. *See* SADOVNIK ET AL., *supra* note 48, at 85–86 (discussing states' actions following *A Nation at Risk* and other reports issued around that time); *see also* STALLINGS, *supra* note 171, at 6.

190. *See* Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988, Pub. L. No. 100-297, § 1001, 102 Stat. 130, 140 (1988) (reauthorizing the ESEA); *see also* Thomas & Brady, *supra* note 165, at 54.

191. Thomas & Brady, *supra* note 165, at 54.

192. *Id.* at 54–55.

193. MARZANO ET AL., *supra* note 68, at 20–21. The Hunter Model consisted of seven elements for a lesson's framework: "[a]nticipatory set," "[o]bjective and purpose," "[i]nput," "[m]odeling," "[c]hecking for understanding," "[g]uided practice," and "[i]ndependent practice." *Id.* at 21.

provided a lesson design model that introduced, taught, and then reinforced teacher growth and learning in a variety of ways.[194] Teachers reviewed the model requirements with administrators and were observed and evaluated for their ability to use the model appropriately; student achievement was presumed if the model was properly implemented.[195] Alternate evaluation models were also proposed in the 1980s; some emphasized individualized career development for teachers, and others proposed different types of evaluation and oversight depending on the teacher's experience, age, and developmental level, but the Hunter model was foremost.[196]

The quest for improved educational quality continued into the 1990s, along with a desire for accountability at all levels of the educational process, as discussed at the 1989 historic governors' education summit that focused on the problems raised by *A Nation at Risk* and set lofty goals for improvement.[197] The 1990s also saw a corresponding increase in individual school accountability levels, along with federal government proposals to increase academic standards. Model teacher standards were developed, and national uniform teacher licensing was proposed.[198] However, federal policymakers continued to propose national educational standards and accompanying testing,[199] although they continued to struggle with how best to achieve both high standards and correspondingly high test scores.[200]

In 1996, *Enhancing Professional Practice: A Framework for Testing*[201]

194. *Id.* at 20.

195. *Id.*

196. *See id.* at 21–22 (discussing alternate evaluation models); Tracy, *supra* note 59, at 324 (noting alternate models and emphasizing the popularity of the Hunter model).

197. MARIS A. VINOVSKIS, NAT'L EDUC. GOALS PANEL, THE ROAD TO CHARLOTTESVILLE: THE 1989 EDUCATION SUMMIT 39–40 (1999), https://govinfo.library.unt.edu/negp/reports/negp30.pdf.

198. *E.g.*, INTERSTATE NEW TEACHER ASSESSMENT & SUPPORT CONSORTIUM, COUNCIL OF CHIEF STATE SCH. OFFICERS, MODEL STANDARDS FOR BEGINNING TEACHER LICENSING, ASSESSMENT, AND DEVELOPMENT: A RESOURCE FOR STATE DIALOGUE (1992), http://programs.ccsso.org/content/pdfs/corestrd.pdf.

199. *See, e.g.*, RAVITCH, NATIONAL STANDARDS, *supra* note 154, at 57–58. In 1990, President George H.W. Bush announced his agreed-upon educational goals as (1) early childhood education; (2) increased high school graduation rates; (3) demonstrated academic and practical competency in 4th, 8th, and 12th grade; (4) "first in the world" status in science and math; and (5) literacy and "skills needed to compete in a global economy." *See id.*

200. *Id.* at 58.

201. CHARLOTTE DANIELSON, ENHANCING PROFESSIONAL PRACTICE: A FRAMEWORK FOR TEACHING (2d ed. 2007).

was published by Charlotte Danielson and, because of its popularity, became the professional standard for teacher evaluation.[202] The Danielson Model defined teacher evaluation as assessment in the areas of "[p]lanning and [p]reparation, the [c]lassroom [e]nvironment, [i]nstruction, and [p]rofessional [r]esponsibilities."[203] The Danielson Model was more comprehensive than previous models and examined 76 different components of effective teaching; it also provided a more extensive ranking system for teachers: "unsatisfactory, basic, proficient, and distinguished."[204]

As education received increasingly more public attention, it became a primary component of political campaigns in an unprecedented way.[205] In the late 1980s and early 1990s, politicians began to push for changes in educational policy; they urged the adoption of uniform educational standards and uniform assessments, along with rewards for high-performing schools.[206] Although these proposals were initially unsuccessful, in 1994 Congress adopted the reauthorization of the ESEA, known as Improving America's Schools Act (IASA).[207] The act required each state to develop academic standards to be measured through annual objective testing; schools that failed to increase student achievement were penalized.[208]

The higher academic standards and annual test measurements included in the IASA became the foundation upon which the No Child Left Behind Act (NCLB) was built.[209] Enacted in 2002, the act not only attempted to provide equal educational quality to poor and minority students, it also attempted to reform schools nationwide by holding schools accountable for student achievement.[210] The act limited its reach to states and local educational agencies (LEAs) and did not yet extend the penalties of failing to meet achievement goals to individual teachers, although holding teachers

202. MARZANO ET AL., *supra* note 68, at 23.

203. *Id.*

204. *Id.* at 24–25.

205. *See* STALLINGS, *supra* note 171.

206. Kenneth Jost, *Revising No Child Left Behind*, 20 CQ RESEARCHER 339, 347 (2010).

207. James E. Ryan, *The Perverse Incentives of the No Child Left Behind Act*, 79 N.Y.U. L. REV. 932, 938 (2004); *see also* Improving America's Schools Act of 1994 (IASA), Pub. L. No. 103-382, 108 Stat. 3518 (codified as amended in scattered sections of 20 U.S.C.).

208. Ryan, *supra* note 207, at 938–39.

209. *Id.* at 939–40.

210. *Id.*; *see also* No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (codified as amended in scattered sections of 20 U.S.C.).

accountable became the next logical step in the process.

III. NCLB AND THE QUEST FOR ACCOUNTABILITY

The twenty-first century saw a sea of change in teacher evaluation structure as "emphasis . . . shifted from supervision to evaluation, as well as from teacher behavior to student achievement."[211] The first stage of the shift was holding schools accountable for student learning. The NCLB integrated the language of accountability into educational practice by offering increased funding to schools complying with federal goals, including an increased emphasis on testing and a pattern of continual improvement.[212] The second stage of the shift was to hold teachers individually accountable for student learning as measured by standardized testing.

The NCLB increased levels of school accountability and was rooted in the notion that higher educational standards, better instruction, new programs, and increased school choice would result in increased learning and test performance.[213] Accountability was imposed through yearly testing and public reporting of test results.[214] Schools that failed to demonstrate annual yearly progress were penalized.[215] However, some of that accountability was illusory. While the NCLB was credited with increasing school accountability, by 2002 when the act was reauthorized and created, 39 states already had such systems in place.[216]

The effects of the NCLB were mixed. While states complied with its requirements in order to secure funding, the drawback was the increased use of high-stakes testing models with performance targets set by the federal government—targets and their associated penalties that could not easily be tailored to the needs of each state.[217] Additionally, the many criteria associated with the act made it difficult to see whether increased testing and

211. MARZANO ET AL., *supra* note 68, at 25.

212. Alyson Klein, *No Child Left Behind: An Overview*, EDUC. WK. (Apr. 10, 2015), http://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html.

213. Kimberly A. Murakami, *Construction and Application of No Child Left Behind Act, Pub. L. No. 107-110, 115 Stat. 1425 (2002) (codified at 20 U.S.C.A. §§ 6301 et seq.)*, 4 A.L.R. Fed. 2d 103, § 2, at 107 (2005).

214. Ryan, *supra* note 207, at 942.

215. *Id.* at 942–43.

216. Eric A. Hanushek & Margaret E. Raymond, *Does Accountability Lead to Improved Student Performance?*, 24 J. POL'Y ANALYSIS & MGMT. 297, 298 (2005).

217. *See id.*

accountability were successful when looking at varied state policies and disparate populations.[218] For example, Eric Hanushek and Margaret Raymond found that the accountability required by the NCLB led to improved test results when implemented along with a penalty and reward system.[219] However, the imposition of accountability did not lead to uniform, improved results for all groups—some subgroups showed improved performance over others.[220]

While there were modest gains made in student growth under the NCLB, disparities for minority and socioeconomically challenged students remained.[221] The NCLB's promise that the student achievement gap would be closed was not borne out by the results. While there were some gains in fourth and eighth grade math, there were no reading gains at all.[222] The NCLB's goal of bringing all students to the "proficiency" level by 2014 was a goal that no state met.[223] As the effort to transform education seemed to become more frustrating and less fruitful, the federal government began to press further for changes to teacher evaluation. The NCLB and its failures—as demonstrated by lack of test score improvement—seemed to highlight the need for greater teacher accountability.

In 2011, to moderate some of the effects of the NCLB, President Barack Obama introduced a waiver system under which states could discontinue compliance with the NCLB in return for adoption of "college- and career-ready standards"[224] and improved teacher evaluations that included student achievement, as measured by objective testing.[225] In

218. *Id.*

219. *Id.*

220. *Id.* ("When we look specifically at the performance of subgroups, we find that Hispanic students gain most from accountability while Blacks gain least.").

221. *Data: Student Achievement in the Era of Accountability*, EDUC. WK. (Dec. 30, 2015), http://www.edweek.org/ew/qc/2016/data-student-achievement-in-the-era-of.html?intc=EW-QC16-TOC.

222. BAKER ET AL., *supra* note 38, at 6.

223. Klein, *supra* note 212.

224. Michele McNeil & Alyson Klein, *Obama Offers Waivers from Key Provisions of NCLB*, EDUC. WK. (Sept. 27, 2011), http://www.edweek.org/ew/articles/2011/09/28/05waiver_ep.h31.html. These standards have been adopted and are called "common core" standards by the states. *See id.*; *see also College- and Career-Ready Standards*, U.S. DEP'T EDUC., http://www.ed.gov/k-12reforms/standards (last visited Nov. 29, 2016).

225. McNeil & Klein, *supra* note 224. In exchange for a waiver, states were required to agree either to adopt common core standards or to be certified as meeting "college-

addition to waiving the NCLB standards in favor of the common core, the Obama Administration furthered its educational goals through Race to the Top (RTTT).[226] RTTT granted states "monetary incentives to reform their educational systems in certain ways,"[227] such as the development of evaluation processes that "improv[e] teacher and principal effectiveness based on performance."[228] Although states were encouraged to develop their own teacher evaluation guidelines, RTTT rewarded those teachers who generated consistently high test scores.[229]

RTTT built on the accountability model introduced by the NCLB and provided financial rewards to states that included increased student learning measures into their evaluation practices.[230] RTTT emphasized incorporating objective data into the evaluation process, a practice many teachers opposed.[231] An accompanying part of RTTT was the passage of new laws that limited the ability of teachers' unions to collectively bargain for evaluation practices and the introduction in several states of integrated models of teacher evaluation: observation, preparation, interaction, and student test data.[232] Of the 25 states that integrated student test scores into teacher evaluations, 20 assigned a 30 to 50 percent value to student growth indicators.[233] RTTT reforms resulted in increased evaluation for all teachers.[234] It also linked student test data to teacher evaluation scores,

and career-ready" educational standards by higher education institutions in their state. *Id.* Teachers were required to demonstrate student progress on state assessments as part of the evaluation process, and the lowest 15 percent of schools were required to be flagged for improvement. *Id.*

226. RACE TO THE TOP, *supra* note 35, at 2.

227. JUDITH LOHMAN, OFFICE OF LEGISLATIVE RESEARCH, NO. 2010-R-0235, COMPARING NO CHILD LEFT BEHIND AND RACE TO THE TOP (2010), https://www.cga.ct.gov/2010/rpt/2010-r-0235.htm.

228. RACE TO THE TOP, *supra* note 35, at 9.

229. For example, the Delaware Talent Cooperative provides financial awards to retain high performing teachers who choose to teach in underperforming schools. EXEC. OFFICE OF THE PRESIDENT & U.S. DEP'T OF EDUC., SETTING THE PACE: EXPANDING OPPORTUNITY FOR AMERICA'S STUDENTS UNDER RACE TO THE TOP 7 (2014), https://www2.ed.gov/programs/racetothetop/setting-the-pace.pdf.

230. *Id.* at 1.

231. Regina Umpstead et al., *An Analysis of State Teacher Evaluation Laws Enacted in Response to the Federal Race to the Top Initiative*, 286 EDUC. LAW REP. 795, 796 (2013).

232. *Id.* at 796–98.

233. *Id.* at 803–804 & tbl.1.

234. *See id.* at 795.

encouraged more precision within the system, and connected evaluations to consequences, such as dismissal.[235]

In 2013, state evaluation systems varied widely because of the impact of RTTT.[236] The experimentation with evaluation systems that RTTT encouraged made it difficult to determine whether some districts had higher quality teachers than others.[237] For example, 11 states required total compliance by all districts with one central plan.[238] Twenty-seven states provided districts with flexibility: in some cases there were multiple models that different school districts could choose to implement; if those districts did not approve the provided models, they had the option to create their own, subject to compliance with state standards.[239] Ten states provided a model with an opt-out provision for alternative systems.[240] Fewer than half of the states that provided plan flexibility to districts required state authorization.[241]

A review of RTTT's impact shows that under its influence schools moved away from their older evaluation models to models that evaluated the teacher's performance in various ways, one of which was to measure the impact teachers had on student achievement. The 2013 Measures of Effective Teaching project report proposed several alternatives to weighting evaluations so that teachers would not "focus too narrowly on a single aspect of effective teaching and neglect its other important aspects."[242] The first proposal was that student test data should be weighted between 33 to 50 percent by districts to achieve the best predictors of teacher competence.[243]

235. *Id.* at 812–13.

236. KATHRYN M. DOHERTY & SANDI JACOBS, NAT'L COUNCIL ON TEACHER QUALITY, STATE OF THE STATES 2013, i–ii (2013), http://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report.

237. *See* INST. OF EDUC. SCIS., NAT'L CTR. FOR EDUC. EVALUATION & REG'L ASSISTANCE, NCEE EVALUATION BRIEF: STATE REQUIREMENTS FOR TEACHER EVALUATION POLICIES PROMOTED BY RACE TO THE TOP 11–12 (2014) http://files.eric.ed.gov/fulltext/ED544794.pdf (discussing state teacher evaluation policies and their alignment with the RTTT program).

238. DOHERTY & JACOBS, *supra* note 236, at 11.

239. *Id.*

240. *Id.*

241. *Id.* at 13.

242. STEVEN CANTRELL & THOMAS J. KANE, BILL & MELINDA GATES FOUND., ENSURING FAIR AND RELIABLE MEASURES OF EFFECTIVE TEACHING 10–11 (2013), http://www.edweek.org/media/17teach-met1.pdf.

243. *Id.* at 15.

However, the study also recommended that weights be evenly dispersed so that teachers were more well-rounded in their teaching efforts and developed skills that targeted other important learning outcomes.[244] This shift to incorporating teacher accountability required states to adapt their previous evaluation models and to determine which statistical method would best demonstrate student test achievement over the course of a year while still emphasizing other important areas of teacher development.[245]

Despite the heightened interest in connecting teaching ability to student learning, and the corresponding administration of exams and compilation of data, teacher performance quality was still a factor only a limited number of times during a teacher's career: at hiring, in determining compensation, when granting tenure, and during discipline and dismissal proceedings.[246] In 2009, regardless of the rating and weighting system used to evaluate teachers, over 94 percent of teachers were still ranked as satisfactory or higher.[247] The way these ranking systems were used led to the conclusion that differences in teacher quality either were not being consistently measured or were not being accurately reported.[248]

By late 2015, 42 states had waivers and were granted flexibility from the NCLB requirements,[249] and on December 10, 2015, President Obama signed the Every Student Succeeds Act (ESSA)—the latest reauthorization of the ESEA.[250] While the ESSA has some elements in common with the NCLB, the act has a greater focus on college and career readiness and has shifted away from measuring student success using uniform national testing—such as annual yearly progress—toward state-driven assessments.[251] This transition from the one-size-fits-all requirements of the NCLB includes multiple forms of student assessment, state-driven standards, intervention and funding for the lowest-performing schools, state determination and

244. *See id.* at 14–15.

245. *See id.*

246. DANIEL WEISBERG ET AL., THE WIDGET EFFECT: OUR NATIONAL FAILURE TO ACKNOWLEDGE AND ACT ON DIFFERENCES IN TEACHER EFFECTIVENESS 4 (2d ed. 2009), http://files.eric.ed.gov/fulltext/ED515656.pdf.

247. *Id.* at 6.

248. *Id.*

249. *ESEA Flexibility*, U.S. DEP'T OF EDUC., http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html (last modified May 12, 2016).

250. *Every Student Succeeds Act (EESA)*, U.S. DEP'T OF EDUC., http://www.ed.gov/essa?src=in (last visited Mar. 15, 2017).

251. *Id.*

creation of evaluation systems, programs to reward effective teachers, commitment to increase the number of STEM teachers, and resources to encourage data-driven systems and creative approaches to education.[252]

IV. VALUE-ADDED MODELS AND TWENTY-FIRST CENTURY TEACHER EVALUATION

As accountability became the accepted tool for motivating change in education, experts began studying whether teacher evaluations could be more fully attached to student achievement as measured by objective testing.[253] William Sanders first pioneered and applied VAM to individual teachers beginning in the mid-1980s to prove that teacher effectiveness could be measured using student test data.[254] Sanders and Robert McLean developed a mixed statistical model to evaluate Tennessee teachers during a period when the state was trying to introduce greater school accountability and wanted to reward high-performing teachers.[255] The model was adopted in 1992, and students were tested each year in grades three to eight; VAM measured students against themselves—revealing whether learning was increasing or flat over time.[256]

There was a growing public interest in seeing concrete evidence of student achievement combined with accountability for such achievement from teachers and administrators.[257] It had long been argued that teacher performance could not be adequately measured using students' standardized testing.[258] On the one hand, students were individuals learning based on their

252. *Id.*

253. David Hill, *He's Got Your Number*, EDUC. WK. TEACHER (May 1, 2000), http://www.edweek.org/tm/articles/2000/05/01/08sanders.h11.html.

254. *Id.*

255. *Id.*

256. *Id.*

257. Further compounding the issue is whether teacher evaluations, including VAM, should be released to the public. Frank G. Barile, Note, *Making Enemies Out of Educators: The Legal and Social Consequences of Disclosing New York City Teacher Data Reports*, 2013 BYU EDUC. & L.J. 125, 128–29 (2013). A New York appellate court ruled that teacher data was accessible to the public through the state's freedom of information law as public agency record. Mulgrew v. Bd. of Educ. of the City Sch. Dist. of New York, 928 N.Y.S.2d 701, 703 (App. Div. 2011). This led the legislature to pass a law that teacher evaluations only be released anonymously or in response to parental request for students currently enrolled in the class. *See* N.Y. EDUC. LAW § 3012(c)(10) (McKinney 2017).

258. *See* BAKER ET AL., *supra* note 38.

own experiences and ability levels; their ability to learn was impacted by their homes, families, and personal circumstances.[259] On the other hand, test performance could indicate whether students were being taught the skills and knowledge they need to succeed in the future.[260]

Using students' objective state test scores to evaluate teacher quality made sense on one level. If a teacher's role is to impart information to students, then the testing instrument adopted by the state to determine whether the student has, in fact, learned the material seems like a fair measurement of what and how well the teacher has taught. Using VAM to isolate teacher effectiveness was also attractive because of the many resources that had already been provided to schools and teachers to improve instruction. Class sizes were smaller; teachers received training, mentoring, and other development opportunities; there were increasing numbers of teachers with graduate degrees; but students still struggled.[261] VAM seemed like a viable option for improving teaching quality; it could filter out variables and show the influence of an individual teacher on student achievement.[262]

In 1992, Erik Hanushek determined that teacher quality "might make as much as a full year's difference in a student's learning growth—a gain of 0.5 grade level equivalents for a student with a low-quality teacher compared with 1.5 for a student with a high-quality teacher."[263] In 2005, another study used VAM to determine that "high quality instruction throughout primary school could substantially offset disadvantages associated with low socioeconomic background."[264] Furthermore, identifying very ineffective teachers could have a positive impact on countless school children over time, and "eliminating the least effective 6–10 percent of teachers would bring student achievement up by 0.5 [standard deviations]."[265]

VAM reveals the impact an individual teacher can have on the life of a child and indicates that states and LEAs should make every effort to

259. *See id.* at 3.

260. *See id.*

261. Hanushek, *Teacher Deselection*, *supra* note 32, at 170.

262. Hill, *supra* note 253.

263. Nancy Protheroe, *Using Student Achievement Measures to Evaluate Teachers*, INFORMED EDUCATOR SERIES, NOV. 14, 2006, at 1, 1.

264. Steven G. Rivkin et al., *Teachers, Schools, and Academic Achievement*, 73 ECONOMETRICA 417, 419 (2005).

265. Hanushek, *Teacher Deselection*, *supra* note 32, at 173.

recruit and train high-performing teachers in minority and lower income schools. "[I]f a student had a good teacher as opposed to an average teacher for four or five years in a row, the increased learning would be sufficient to close entirely the average gap between a typical low income student . . . and the average student . . . ."[266] The impact of this is equally pronounced in the negative: "[T]he least effective 5 percent of teachers see gains that are at best two-thirds of a grade-level equivalent."[267] Some gains are lower than that: "The bottom 1 percent of teachers achieve no more than one-half of a grade-level equivalent in annual gains."[268] Students who are placed with good teachers are advantaged over their counterparts who are placed with poor teachers, and some of the students may never get the gains needed over time should they be taught with overall poor teacher quality in multiple years.[269]

In 2011, Raj Chetty, John Friedman, and Jonah Rockoff evaluated the accuracy of VAM methods using a data set of 2.5 million students who were tested in math and English between 1989 and 2009.[270] The test score data was matched to other family data available through the schools and tax records.[271] The study found that VAM models using such controls were accurate and "implie[d] that improvements in teacher quality [could] raise students' test scores significantly."[272] Highly effective teachers were found to have a long-term measurable impact on student success, including lowering chances of teen pregnancy and increasing the following: (1) the likelihood of college attendance, (2) the chances that the student would attend a better ranked college, (3) the likelihood of greater lifetime income, (4) the likelihood of long-term savings, and (5) the likelihood that the student would live in a desirable neighborhood.[273] The study found that while student recall of content taught by the teacher gradually faded out, the long-term impact of a good teacher went beyond test material.[274] While the study acknowledged that improving test scores is not the final goal of education, it

266. *Id.* at 172.
267. *Id.*
268. *Id.*
269. *Id.*
270. Chetty et al., *Measuring the Impacts of Teachers I*, *supra* note 17, at 2594.
271. *Id.*
272. *Id.* at 2630.
273. Raj Chetty et al., *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood*, 104 AM. ECON. REV. 2633, 2634 (2014) [hereinafter Chetty et al., *Measuring the Impacts of Teachers II*].
274. *Id.* at 2635.

recommended that VAM be used as one measure of teaching quality.[275]

The benefit of VAM is its ability to measure the impact an individual teacher can have on the educational experience of a child. VAM does not necessarily require a constant rate of demonstrated student growth to show teacher competence; it accounts for the fact that students may not learn at a consistent rate.[276] Additionally, the benefit of VAM over other accountability measures is that it does not measure the school at one point in time to determine rewards or penalties based on that one test score.[277] Instead, VAM looks at the growth students make from one year to the next, which eliminates some concern about academic intangibles such as family life, parent involvement, poverty level, and school choice.[278]

VAM is not without its detractors, who argue that student achievement is affected by factors beyond individual teacher quality that are impossible to factor out. One such argument addresses the difficulty in separating teacher quality from student quality given the problem of student assignment into classes; students are rarely "randomly sorted" and teachers are rarely "randomly assigned."[279] Instead, parents who are well-connected at the school often lobby to place their children in classrooms and with teachers who are perceived as the "best" teachers for that grade level.[280] The opposite is true as well; teachers frequently have input into classroom placement, too.[281] This means that,

> [S]tudents assigned to a particular teacher may not be representative of the general student population with respect to their level and rate of growth in achievement, parental support, motivation, study habits, interpersonal dynamics and other relevant characteristics. It is very difficult for the statistical machinery to disentangle these intrinsic student differences from true differences in teacher effectiveness.[282]

275. Chetty et al., *Measuring the Impacts of Teachers I*, *supra* note 17, at 2631.

276. *See* HENRY I. BRAUN, USING STUDENT PROGRESS TO EVALUATE TEACHERS: A PRIMER ON VALUE-ADDED MODELS 6 (2005), https://www.ets.org/Media/Research/pdf/PICVAM.pdf.

277. Rivkin et al., *supra* note 264, at 418–19.

278. *Id.*

279. BRAUN, *supra* note 276, at 3.

280. *Id.* at 7.

281. *Id.*

282. *Id.* at 3.

Given the assignment problem, it can be difficult to ascertain which factor drives student success—excellent teachers or excellent students. After controlling for these factors, Steven Rivkin, Eric Hanushek, and John Kain determined that the assignment problem could be controlled for.[283] The difference in student learning for students with a "good teacher" in math rather than an "average teacher" was "0.11 [standard deviations] of student achievement."[284]

Other frequent teacher concerns with the use of VAM methods were addressed by Jason Millman, who wrote about the lack of control teachers have over external factors that impact student success:

> The single most frequent criticism of any attempt to determine a teacher's effectiveness by measuring student learning is that factors beyond the teacher's control affect the amount students learn. These factors range from those at the student level (e.g., student ability) to those at the classroom (e.g., class size) and school and community levels (e.g., district wealth).[285]

While most policymakers agree that a classroom teacher has the greatest opportunity to influence a child at school, some studies indicate that there are other, greater influences on childhood learning. For example, one study attributes only 9 percent of a student's success to the classroom teacher and 60 percent to external influences.[286]

VAM can be used to measure teacher quality generally and to target specific areas that need improvement for both students and teachers.[287] However, state testing may not be up to the task of fully assessing the many goals of the K–12 educational system since "[t]he subject matter knowledge, the learning skills, the testing formats, and the noncognitive outcomes that are targeted by the reformers/educators are greater than the information accessible by testing."[288]

283. Rivkin et al., *supra* note 264, at 418.

284. Hanushek, *Teacher Deselection*, *supra* note 32, at 171.

285. Jason Millman, *How Do I Judge Thee? Let Me Count the Ways* in GRADING TEACHERS, GRADING SCHOOLS: IS STUDENT ACHIEVEMENT A VALID EVALUATION MEASURE 243, 244 (Jason Millman ed., 1997).

286. Dan D. Goldhaber et al., *A Three-Way Error Components Analysis of Educational Productivity*, 7 EDUC. ECON. 199, 206–07 & tbl.3 (1999).

287. Protheroe, *supra* note 263, at 9.

288. Millman, *supra* note 285, at 245, *quoted in* Protheroe, *supra* note 263, at 7.

Since VAM compares students to their own prior performance, difficulties can also arise regarding incomplete data.[289] The U.S. population is very mobile; students frequently move from district to district.[290] Because students move and transfer schools from year to year, their personal learning is difficult to measure consistently since the data is located in different areas.[291] When students move during the school year, it is difficult to determine which of two (or three) teachers is responsible for the student's achievement.[292]

There is also disagreement surrounding whether VAM can consistently identify strong and weak teachers. For example,

> One study found that across five large urban districts, among teachers who were ranked in the top 20 [percent] of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40 [percent]. Another found that teachers' effectiveness ratings in one year could only predict from 4 . . . to 16 [percent] of the variation in such ratings in the following year. Thus, a teacher who appears to be very ineffective in one year might have a dramatically different result the following year. The same dramatic fluctuations were found for teachers ranked at the bottom in the first year of analysis. This runs counter to most people's notions that the true quality of a teacher is likely to change very little over time and raises questions about whether what is measured is largely a "teacher effect" or the effect of a wide variety of other factors.[293]

Thus, VAM may be successful when used to identify failing and effective teachers on the extreme ends of the spectrum, but it may be more difficult to assess the performance of teachers who "fall somewhere in the middle range."[294] VAM may work far better at assessing schools as a whole rather than individual teachers.[295]

289. Protheroe, *supra* note 263, at 4–6.

290. *See* NAT'L ASS'N OF STATE BDS. OF EDUC., EVALUATING VALUE-ADDED: FINDINGS AND RECOMMENDATIONS FROM THE NASBE STUDY GROUP ON VALUE ADDED ASSESSMENT 21 (2005) [hereinafter NASBE, EVALUATING VALUE-ADDED].

291. ERIC A. HANUSHEK ET AL., DOES IT MATTER HOW WE JUDGE SCHOOL QUALITY? 9 (2004).

292. NASBE, EVALUATING VALUE ADDED, *supra* note 290.

293. BAKER ET AL., *supra* note 38, at 2.

294. NASBE, EVALUATING VALUE ADDED, *supra* note 290.

295. *Id.* at 21–22.

When applying VAM to teacher evaluation, it is notable that state tests are minimum-competency tools measuring only mastery of grade-level standards, and the results are tied to a student's entry-level knowledge base.[296] Thus,

> [T]he additional gain in test scores resulting from a substantial improvement in the quality of instruction may be quite sizeable for a student who begins at the lower end of the skill distribution and for whom the test covers much of the knowledge gained by virtue of any higher teacher quality. On the other hand, a student higher up the initial skill distribution may answer most of the questions correctly even if taught by a quite low quality teacher. Better instructional quality may translate into only a few additional correct answers if the test does not concentrate on or cover the additional knowledge generated for this student by the superior instruction.[297]

Alternatively, students who are academically disadvantaged and begin the year behind academically, and those who have achieved learning growth in lower grade-level areas not scored by their present grade exam, will not score well on the test, as they have not achieved grade-level work.[298] This means that any growth that has occurred cannot be measured with the testing instrument for that year.[299] This high-achiever, low-achiever conflict can create multiple years of unreliable test data.[300]

In measuring whether students have met state grade-level standards, it is important to recognize that "'[m]inimum competency' examinations . . . fall short of what is needed, as the 'minimum' tends to become the 'maximum,' thus lowering educational standards for all."[301] The heavy weighting of test scores in teacher evaluations may encourage teaching to the test, rather than encouraging teachers to offer more varied learning experiences. Also, if the focus on testing and learning remains rooted in the lower grades as it has in the past, it will not show the whole picture of student learning growth nor where weaknesses occur in later years. Strong

296. *E.g.*, Eric A. Hanushek et al., *The Market for Teacher Quality* 6–7 (Nat'l Bureau of Econ. Research, Working Paper No. 11154, 2005), http://www.nber.org/papers/w11154.pdf (discussing the Texas VAM standards).

297. *Id.* at 7.

298. HAERTEL, *supra* note 30, at 16.

299. *Id.* at 8.

300. *Id.* at 8–9.

301. *A Nation at Risk*, *supra* note 175, at 121.

performance on testing in the early grades may not have a strong correlation to upper grade-level student success,[302] but this will be difficult to measure unless current testing structures change.

VAM models are also limited in what they reveal about improving teaching quality. VAM can tell a teacher whether students are consistently demonstrating academic achievement, but cannot isolate what specifically the teacher needs to do to improve teaching effectiveness. While the teacher may know that a student did not understand a basic concept, she may not be able to isolate what she should improve in her teaching methods to cement that concept. A teacher may also teach well without knowing exactly how she got such positive results, since "merely describing the product (what students know and can do) provides scant information on what the teacher did or should have done to yield better results."[303]

Another area of VAM weakness is in potential test manipulation. VAM may drive schools and teachers to discourage students likely to perform poorly from taking exams.[304] Test scores can also be improved by teaching less content and narrowing the academic focus to what will be on the test; students will end up learning less but looking more intelligent.[305] On the other hand, VAM can also improve test scores by improving teaching quality, changing the curriculum, and providing additional supports for students.[306] However, the temptation may be to take the easier route.[307]

VAM measurement can also be challenging when adopting new state standards. To hold teachers accountable and accurately measure student achievement, testing instruments must be high-quality and error-free or the purpose of using the test as a tool for evaluation is lost.[308] Once the test becomes the measurement of teacher success, it will have the effect of driving teachers toward emphasizing those content areas tested to the

302. Student growth measured from grade four to grade eight did not correlate to continued learning growth at age 17. Eric A. Hanushek, *The Importance of School Quality*, *in* OUR SCHOOLS AND OUR FUTURE . . . ARE WE STILL AT RISK? 161–62 & fig.5 (Paul E. Peterson ed., 2003).

303. Millman, *supra* note 285, at 247.

304. Hanushek & Raymond, *supra* note 216, at 300–01, 303.

305. *Id.* at 300–01.

306. *Id.* at 303.

307. *See id.* (noting that removing poor student scores from school scores is easier than improving teaching and providing student-help programs).

308. *See* Protheroe, *supra* note 263, at 7–8.

exclusion of others.[309] If the test is low-quality, nothing will be gained by the child and content knowledge will be reduced, rather than increased.[310]

The testing issues discussed above illustrate how VAM cannot be too heavily weighted in deciding whether teachers with the highest test scores are "good" teachers. VAM may only indicate that a teacher is better at teaching to the test, while another teacher may in fact be providing a more well-rounded educational experience for students, some of which will not be measured by the test.[311] VAM also fails to hold equally accountable other teachers who are impacting student achievement, but whose subjects are not tested by state exams.[312] This places an unfair burden on teachers teaching core subjects and has the potential to drive them to teaching in untested subject areas. Adapting the VAM framework to high schools is also problematic: while the elementary school system places the responsibility for learning on one teacher, there are multiple, desirable high school subjects that are tested by college board exams, rather than by state achievement testing.[313] VAM also cannot fully address cross-subject learning transfers;[314] for example, multiple school subjects may influence a student's reading score beyond what the student has learned in language arts.[315]

Implementation is another VAM challenge. It may be difficult for local school districts to implement with their current employees since complex statistical analysis is required.[316] In addition, VAM necessarily requires access to multiple years of accurate data, which schools must collect and save.[317] The increasing data and recordkeeping needed to utilize VAM, eliminate bias, and ensure its success[318] are dependent on the district's hiring well-qualified staff or external firms to do the work.

309. *See id.* at 8.

310. *Id.*

311. *Id.* at 6.

312. *See* NASBE, EVALUATING VALUE ADDED, *supra* note 290.

313. DAN GOLDHABER ET AL., CT. FOR EDUC. DATA & RESEARCH, TEACHER VALUE-ADDED AT THE HIGH SCHOOL LEVEL: DIFFERENT MODELS, DIFFERENT ANSWERS? 5 (2010), http://www.cedr.us/papers/working/CEDR%20WP%202011-4%20Value-added%20Assessment%20(10-19-2011).pdf.

314. *See id.* at 19 ("With the exception of Geometry, the hypothesis of no cross-subject teacher effect is consistently rejected at the 95 percent confidence level.").

315. *See* NASBE, EVALUATING VALUE ADDED, *supra* note 290, at 20–21.

316. *See* HAERTEL, *supra* note 30, at 10 (describing VA models).

317. *See id.*

318. *See* Protheroe, *supra* note 263, at 6.

Moreover, VAM may have a disparate impact on students enrolled in schools that serve low-income and minority students.[319] The concern is that teachers will be driven away from teaching in poverty-level schools because they will not be consistently able to demonstrate student achievement given the make-up of the population.[320] The avoidance effect may also limit the number of teachers willing to take on the challenge of preparing special education and English Language Learner (ELL) students for testing.[321] Teachers may opt out of teaching such students or divert those students into untested subjects to avoid the possibility that the teacher will not be able to show sustained learning growth. VAM may also have adverse consequences on low-performing schools.[322] Students at those schools may be placed in the unenviable position of having a curriculum imposed on them that teaches solely to the test, denying them other rich learning opportunities that their counterparts in wealthier neighborhoods might receive.

Just as RTTT removed the prohibition on connecting teacher performance evaluations to test scores,[323] further incorporation of VAM as the primary model for evaluation lessens union bargaining power and prior agreements and policies, such as "last in, first out."[324] It may also diminish the union's power to intercede in employment actions.[325] Since any state evaluation scheme will be adjudicated using a rational basis test, the state needs only show a legitimate government interest in the exercise of its power to prevail.[326] Many hard-won bargaining agreements may vanish as a result of VAM.

Privacy is also a concern for teachers evaluated using VAM. The

319. Rebecca Dizon-Ross, How Do School Accountability Reforms Affect Teachers? Evidence from New York City 1 (Dec. 3, 2014) (unpublished manuscript) (on file with Harvard Library), http://faculty.chicagobooth.edu/rebecca.dizon-ross/research/papers/accountabilityTeachers.pdf.

320. *Id.*; *see also* NASBE, EVALUATING VALUE-ADDED, *supra* note 290, at 26 (discussing how the odds are stacked against high-poverty schools).

321. *See* BAKER ET AL., *supra* note 38, at 3–4.

322. *See* Dizon-Ross, *supra* note 319.

323. Umpstead et al., *supra* note 231, at 796–97.

324. *See* William S. Koski, *Teacher Collective Bargaining, Teacher Quality, and the Teacher Quality Gap: Toward a Policy Analytic Framework*, 6 HARV. L. & POL'Y REV. 67, 86–87 (2012).

325. *See id.*

326. *See* Regina Umpstead et al., *The New State of Teacher Evaluation and Employment Laws: An Analysis of Legal Actions and Trends*, 332 EDUC. L. REP. 577, 585 (2015) (citing Cook v. Bennett, 792 F.3d 1294, 1300 (11th Cir. 2015)).

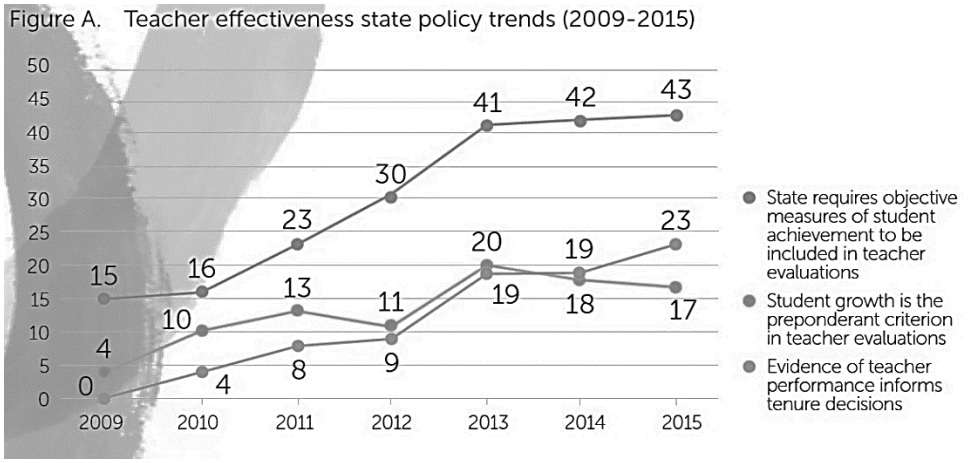
availability of data and individual teacher scoring raises the question of what and how much information should be provided to parents and communities.[327] Privacy concerns may eventually drive teachers away should individual teacher data not be kept secure. These concerns have to be balanced against the public's interest in knowing where high-quality teachers are located. Currently only 13 states require that districts disclose aggregate teacher evaluation data so that parents can make decisions such as home buying and school choice,[328] but this is likely to change in the future. Use of VAM to reveal information about individual teachers must be fair and uniform. To achieve this, VAM must be limited to measuring student achievement rather than being used to compare teacher test scores for each grade to see who is the "best" teacher and who is the "worst."

V. CURRENT STATE EVALUATION MODELS

As a result of the call for greater teacher and school accountability, the stringent requirements of the NCLB, the increased popularity of VAM as a component of evaluation, and the incentives of RTTT, states have shifted teacher evaluation systems radically. Instead of planning, observation, and feedback models (such as the Danielson Model), states are allocating increasing percentages of a teacher's evaluation to student-test data. The table below illustrates the changing landscape of teacher evaluation:

327. *See, e.g.*, Mulgrew v. Bd. of Educ. of the City Sch. Dist. of New York, 919 N.Y.S.2d 786, 787–89 (Sup. Ct.), *aff'd*, 928 N.Y.S.2d 701 (App. Div. 2011).

328. KATHRYN M. DOHERTY & SANDI JACOBS, NAT'L COUNCIL ON TEACHER QUALITY, STATE OF THE STATES 2015, at 32 (2015), http://www.nctq.org/dmsView/StateofStates2015 [hereinafter DOHERTY & JACOBS, STATE OF THE STATES 2015]. The states are Arkansas, Colorado, Florida, Illinois, Indiana, Louisiana, Massachusetts, Michigan, Missouri, New York, North Carolina, Ohio, and Pennsylvania. *Id.*



329

Initially, VAM assessments were a minimal part of evaluation systems and played no role in tenure decisions.[330] However, over the past seven years, states have increasingly incorporated student learning growth into teacher evaluations.[331] While the impact of VAM on tenure has grown far more slowly, it appears to be an increasing trend.[332] In addition to the rising use of VAM, the number of states requiring annual evaluations for all teachers, rather than just for those with probationary status, has also grown (from 15 in 2009 to 27 in 2015); 45 states require yearly review of probationary teachers.[333] The emphasis on locating ineffective teachers for remediation or dismissal is growing.

Currently the development and application of teacher evaluation models varies widely from state to state.[334] There is additional variety in the content and instruments used to measure student achievement from year to year.[335] The quantity of evaluations varies as well. Some states provide

329. Chart reprinted with permission of the National Council on Teacher Quality. *Id.* at i fig.A.

330. *See id.*

331. *See id.* at 1.

332. *See id.* at 2 fig.1.

333. *Id.* at ii.

334. *See, e.g.*, *id.* at 7 fig.5 (summarizing the key teacher evaluation requirements in all 50 states).

335. *See, e.g.*, *id.* at 11 fig.8 (summarizing schoolwide student growth measures across various states).

annual teacher observation only for new teachers, whereas older, more experienced teachers are either not evaluated or are evaluated on a rolling basis.[336]

VAM makes up an increasing portion of a teacher's overall evaluation score and, in some states, is tied to compensation schemes.[337] However, the question of whether and how to connect teacher salary to teacher performance continues to be a difficult issue for states.[338] Past studies of school accountability showed that any system attempting to impose accountability must have a penalty or reward attached to it for efficacy.[339] As Eric Hanushek and Margaret Raymond showed in their analysis of school achievement growth under the NCLB, reporting data alone was ineffective absent consequences.[340] It is difficult to tell how this might apply to individual teacher quality, as states have tied compensation to student achievement. Some states have added to this difficulty by delaying the consequences of poor evaluation scores, while others have not yet adopted this model for a variety of reasons, such as prior collective bargaining agreements.[341]

Another piece of constructing effective teacher evaluation models is whether and how much VAM should affect low-performing teachers. The current practical effects of low evaluations on teachers are shown below:

- In 29 states teachers who rank low on the evaluation scale must follow a prescribed improvement plan.[342]
- In 24 states teachers with low evaluations may face dismissal.[343]
- In 19 states teacher tenure is affected by evaluations.[344]
- In 15 states teacher lay-offs are determined using evaluations.[345]
- In 2 states teacher evaluations affect licensure.[346]

336. *See, e.g.*, *id.* at 15–16 fig.12.

337. *Id.* at v–vi (listing states that "directly tie teacher compensation to teacher evaluation results").

338. *Id.* at 33.

339. *E.g.*, Hanushek & Raymond, *supra* note 216.

340. *Id.* at 321.

341. *See* DOHERTY & JACOBS, STATE OF THE STATES 2015, *supra* note 328, at iii.

342. *Id.* at 30.

343. *Id.*

344. *Id.*

345. *Id.*

346. *Id.*

Despite the incorporation of VAM into teacher evaluation models, it is unclear whether they are getting the desired result—more effective teaching.[347] For example, a 2015 study found that even with more stringent evaluation practices put in place by the state, most teachers were given the rating of "effective" or better by administrators.[348] This suggests that the evaluations are not being properly implemented, and may suggest that multiple evaluators are necessary to achieve true objectivity.[349]

Even though states have incorporated more stringent teacher evaluation systems and connected those systems to compensation, the teacher accountability problem remains as school systems and their quality and rigor can vary widely from state to state. In the 1990s, an educational research study found that 99.8 percent of teacher evaluations were done through direct classroom observation.[350] We now have "[a] more balanced approach to teacher evaluation . . . [that] involve[s] an assessment of the *act* of teaching as well as the *results* of teaching."[351] Nevertheless, VAM proponents recommend "a substantial part of [teacher] evaluation, but not the entirety—perhaps one-third to two-thirds of a total score—should be tied to student test scores in one form or another."[352] This will require schools to continue to track student test scores over time,[353] while maintaining or refining their pre-VAM evaluation systems to be more responsive to student achievement.

VI. POLICY RECOMMENDATIONS FOR CREATING EFFECTIVE TEACHER EVALUATION MODELS

The controversy surrounding teacher evaluation and retention may be based on a public perception that is more imagined than real. In reviewing the number and types of legal actions taken by school districts against teachers, Perry Zirkel found that despite widespread public belief that

347. NASBE, EVALUATING VALUE-ADDED, *supra* note 290, at 5.

348. DOHERTY & JACOBS, STATE OF THE STATES 2015, *supra* 328, at iii.

349. *Id.* at iii, 14 (noting that using multiple observations and raters improves teacher perception and provides for greater differentiation).

350. PAMELA D. TUCKER & JAMES H. STRONGE, LINKING TEACHER EVALUATION AND STUDENT LEARNING 7 (2005).

351. *Id.*

352. Protheroe, *supra* note 263, at 4 (alteration in original) (quoting DOUGLAS O. STAIGER ET AL., BROOKINGS INST., IDENTIFYING EFFECTIVE TEACHERS USING PERFORMANCE ON THE JOB 19 (2006)).

353. *Id.*

employment cases are more likely to favor the teacher, the district is far more likely to be the prevailing party—especially if the district has provided due process rights.[354] This is true even when the district has not stringently followed its own policies.[355] Thus, the real issues with retaining low-performing teachers is likely to be that they are not clearly identified or that the district chooses not to take action, as seen above in *Doe v. Antioch*.[356]

While using VAM as part of overall teacher evaluation seems to be the current favored method, it is not the only measure of teacher quality. When determining when and how VAM should be used, it is important to consider (1) the varied purposes of education;[357] (2) the need to provide higher quality education to children in lower socioeconomic areas;[358] (3) the importance of state and local control over education;[359] (4) the professionalization of teachers;[360] (5) the potential impact stringent evaluations may have on teacher supply;[361] (6) the need for changes to teacher education and post-hiring training programs;[362] (7) the shared accountability for student achievement between teachers and administrators;[363] and (8) the significant differences between measuring job performance in businesses versus classrooms.[364]

A. *The Purposes of Education*

Schools serve multiple purposes, a fact which is often overlooked by those calling for increased testing and corresponding accountability. While the "specific purposes of schooling are intellectual, political, social, and economic," the "intellectual purpose" is "to teach basic cognitive skills such as reading, writing and mathematics" that leads to "higher order thinking skills such as analysis, evaluation, and synthesis."[365] This is the purpose we

354. Perry A. Zirkel, *Case Law for Performance Evaluation of Public School Professional Personnel: An Update*, 314 EDUC. L. REP. 1, 1–3 (2015).

355. *Id.* at 7.

356. *See supra* Part I.

357. *See infra* Part VI.A.

358. *See infra* Part VI.B.

359. *See infra* Part VI.C.

360. *See infra* Part VI.D.

361. *See infra* Part VI.E.

362. *See infra* Part VI.F.

363. *See infra* Part VI.G.

364. *See infra* Part VI.H.

365. SADOVNIK ET AL., *supra* note 48, at 20.

typically focus on when thinking of education and is the main thrust of current evaluation policies.[366] The political purpose of education is "to prepare citizens . . . [to] participate in [the] political order . . . and . . . teach children the basic laws of society," while the social purpose of education is "to socialize children into the various roles, behaviors, and values of . . . society," which provides stability.[367] Finally, the "economic purposes of schooling [are] to prepare students for their later occupational roles" and are relevant for their effect on the economic future of the nation.[368]

Allocating too high of a percentage of the teacher evaluation score to VAM means weighting intellectual development more heavily than other purposes, which may not be advantageous. Weighting one component more heavily directs resources away from the other important purposes that then might not be fully provided to students educated in the current system.

Next, while we tend to view core curriculum subjects as drawing on a uniform and accepted body of knowledge, the reality is quite different.[369] "The practice of selecting school knowledge involves a process of reconceptualizing and reformulating knowledge" "through a process of selective appropriation, relocation, and refocus that reorders the school subjects."[370] The refined curriculum departs from the original concept after this process.[371] Additionally, the soft skills that are a part of classroom management, such as the ability to work in groups or to engage in reflective learning, are difficult to measure.[372] The ability to teach in these ways may be desirable and helpful to students, but such methods may have little correlation to test performance.

The move toward VAM and high-stakes testing may cause states to ignore other potentially valuable evaluation measures, such as student–parent surveys, lesson plan reviews, teacher self-assessments, measures of

366. NAT'L EDUC. ASS'N, TEACHER ASSESSMENT AND EVALUATION: THE NATIONAL EDUCATION ASSOCIATION'S FRAMEWORK FOR TRANSFORMING EDUCATION SYSTEMS TO SUPPORT EFFECTIVE TEACHING AND IMPROVE STUDENT LEARNING 2 (2010), http://www.nea.org/assets/docs/HE/TeachrAssmntWhtPaperTransform10_2.pdf [hereinafter NAT'L EDUC. ASS'N TEACHER ASSESSMENT AND EVALUATION].

367. SADOVNIK ET AL., *supra* note 48, at 20.

368. *Id.*

369. TUCKER & STRONGE, *supra* note 350, at 101 (discussing difference in curriculum relevant to standardized testing).

370. Popkewitz, *supra* note 75, at 10.

371. *Id.*

372. *See id.*

professional learning, student samples, teacher portfolios, and teacher observation by principal or peers. These may be valuable sources of information for teacher improvement and growth.[373] For example, student surveys combined with VAM and classroom observation provide a more complete overview of effective teaching and its required components than any one measure alone.[374]

The once-common end-of-the-year comprehensive exam needs to be re-evaluated as a tool for assessing teacher quality and student achievement. Evaluation systems should measure students at intervals over time. This allows teachers time to review and reinforce subject matter that has not been fully acquired and gives them time to adapt teaching methods throughout the year to improve effectiveness.

Pre-testing at beginning of the year and post-testing at year's end provide better information about student growth and have the additional benefit of showing student retention of concepts from year to year. However, any VAM assessment tied to testing must be sensitive to the impact of new state standards and the accuracy of new tests. In a time of shifting standards, there may need to be frequent adjustments to testing instruments since the likelihood of error is high.

B. *The Provision of High-Quality Education to Disadvantaged Students*

Greater accountability has been identified as the bridge to the learning gap between middle-class and low-income schools. However, accountability alone will not be enough to bridge the gap for some minority students. A 2005 study found that the school accountability does not lessen the academic achievement gap for all students.[375] Although accountability policies improve student learning over time, there appears to be a greater benefit to students at the top of the scale than to students at the bottom of the scale.[376] Thus, teacher accountability for student achievement is not a complete solution to the teaching–learning dichotomy. States must develop legislation that reflects this to achieve more equal outcomes for all groups.

While the rhetoric surrounding education references a broken system,

373. *See* CANTRELL & KANE, *supra* note 242, at 20–21.

374. KATA MIHALY ET AL., A COMPOSITE ESTIMATOR OF EFFECTIVE TEACHING 38–41 (2013), http://www.rand.org/pubs/external_publications/EP50155.html.

375. *See* Hanushek & Raymond, *supra* note 216, at 310–14.

376. *Id.* (discussing data evidencing lower growth for black and Hispanic students, as compared with white students).

the current educational model works fairly well for middle-class and upper-class students but does not provide the same opportunity for growth for lower-income students.[377] One reason may be the greater availability of enrichment programs.[378] Students placed in superior academic programs tend to perform better.[379] Regrettably, decisions about whether to place students in programs like Gifted and Talented Education tend to have more to do with race and social class.[380] Students in better programs get access to better teaching, better enrichment activities, and better technology.[381] Students in higher socioeconomic groups also suffer fewer learning losses during summer breaks; typically students average about one month's worth of losses in reading skills over the course of a summer.[382] However, middle-income students are more likely to gain reading skills over the summer, while low-income students are more likely to lose such skills.[383] Given the many intangible benefits that accrue to higher-income students, it is not surprising that "high absolute scores on assessments such as the SAT are best predicted by family income."[384]

Using VAM as a majority indicator of teacher quality ignores three demographic variables that strongly influence childhood education: minority status, socioeconomic status, and parent educational level.[385] While education is viewed by many as the way to escape poverty, minority and low-income students can become trapped in an educational system that does not favor them.[386] For example, students with less educated parents typically do not have the same opportunity for academic success as their more fortunate

377. *See* BAKER ET AL., *supra* note 38, at 3.

378. *Id.*

379. *See* SADOVNIK ET AL., *supra* note 48, at 133, 451.

380. *Id.* at 133.

381. *Id.*

382. BAKER ET AL., *supra* note 38, at 3.

383. *Id.*

384. TED HERSHBERG, VALUE-ADDED ASSESSMENT AND SYSTEMIC REFORM: A RESPONSE TO AMERICA'S HUMAN CAPITAL DEVELOPMENT CHALLENGE 5 (2005), http://www.cgp.upenn.edu/pdf/aspen.pdf (originally presented at Aspen Institute's Congressional Institute on the Challenge of Education Reform: Standards, Accountability, Resources and Policy).

385. *See* Jung-Sook Lee & Natasha K. Bowen, *Parent Involvement, Cultural Capital, and the Achievement Gap Among Elementary School Children*, 43 AM. EDUC. RES. J. 193, 204 (2006).

386. *See id.* at 193–94.

counterparts in middle-class schools.[387] One way for students in lower-income schools to succeed at greater rates may be involving their parents in the educational process to the greatest extent possible.[388]

However, parental involvement likely cannot remove all the potential learning barriers students face.[389] For example, studies show that while parent involvement has a positive impact on academic success for all groups, higher income, Caucasian students receive a greater benefit than minority and socioeconomically challenged students.[390] Because of discomfort, work schedules, and lack of familiarity with the educational system, parents without higher education may be unable to spend significant time at school and may not receive important information about what is required to help their children succeed in the current and future school years.[391] In fact, some studies suggest that assigning homework increases the learning gap between wealthy and poor students as poorer parents have less time to help their children with schoolwork at home.[392] Furthermore, it may appear to teachers and administrators that these parents are not vested in the students' success; this may have a trickle-down effect on the teachers' perception of the students' abilities and thus become a self-fulfilling prophecy.[393]

Students in schools with lower socioeconomic indicators need more support and greater enrichment than their wealthier counterparts. To use the same educational system for all ignores the challenges faced by children living in poverty. It also ignores other reasons why these students might not show the same levels of achievement: dangerous neighborhoods,[394] lack of food and sleep,[395] and lack of access to medical, hearing, vision, and dental care[396] are just some possibilities. More creative solutions are needed for

387. *Id.* at 204.

388. *See id.* at 206.

389. *See id.* 209.

390. *E.g.*, *id.*

391. *See id.* at 214.

392. *Cf., e.g.*, *id*. at 212. *But see id*. at 213.

393. *Id.* at 210.

394. Geoffrey T. Woodtke et al., *Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation*, 76 AM. SOC. REV. 713, 713 (2011).

395. *Good Eating and Sleeping Habits Help Kids Succeed in School*, SCIENCEDAILY (July 26, 2013), https://www.sciencedaily.com/releases/2013/07/130726191525.htm.

396. *E.g.*, CHARLES E. BASCH, HEALTHIER STUDENTS ARE BETTER LEARNERS: A MISSING LINK IN SCHOOL REFORMS TO CLOSE THE ACHIEVEMENT GAP 12 (Campaign for Educ. Equity, EQUITY MATTERS: Research Review No. 6, 2010),

children living in these conditions, such as schools that provide all of the above essentials under one roof. Smaller LEAs may also allow districts to be more responsive to the plight of children in lower socioeconomic circumstances as their oversight would be limited to fewer schools, rather than requiring a focus on many schools, all of which have different needs.

The fact that middle-class students have better external opportunities to help them succeed academically results in their better performance on standardized tests and better access to higher education and opportunity in the future. Our goal should be to create an evaluation system that encourages quality teaching and creates a K–12 system without barriers to success; the schools children attend should not limit their futures.

C. *The Shared Accountability of States, LEAs, Schools, and Teachers*

While the federal government has increasingly used its financial power to encourage accountability and lobby for change to evaluation policies, the primary control over such policies must be retained by state and local governments.[397] States and LEAs need flexibility to determine and meet the needs of their various student populations.[398] For example, the needs of individual children within each LEA may vary from school to school: districts with large populations of ELL students may have different goals than other districts nearby; rural school districts may have fewer hiring options than larger urban areas, which could make VAM a barrier to retaining personnel.

This is not to say that states should disregard evaluation factors when making personnel decisions. Instead, decisions regarding retention and dismissal should be kept in the hands of the districts where responses to community concern can be swift. Local accountability allows greater participation in the process for parents and students who have better access to local administrators and a greater opportunity to be heard in a smaller arena than they would on a state or national level.

States are the ultimate arbiters of what their teacher tenure and evaluation models should look like. While the federal government can influence such decisions through a penalty or reward system, each state must determine its goals for content knowledge, social skills, and life skills that

https://www.cde.state.co.us/sites/default/files/documents/healthandwellness/download/healthier_students_are_better_learners.pdf.

397. *See supra* Parts II–III.

398. *See* Woodtke et al., *supra* note 394.

they plan to assess through teacher evaluations.[399] These goals should also be used in allocating teacher time so that teaching is primary and administrative workload is not too burdensome. Thus, any evaluation scheme should keep most school resources earmarked for teaching since it affects individual children in a way that data collection does not. Any evaluation model must also be cost-effective and efficient. This includes being thorough in evaluations and complying with state and federal laws to avoid costly lawsuits filed to force changes to educational policies—such as those currently pending in many large markets. These suits divert funds away from their educational purpose and are a drain on time and resources.

As states transition toward the new college and career readiness standards under ESSA, they must again change their evaluation processes to incorporate new goals.[400] Changes must be planned and deliberate since states are still trying to best determine how to assess student learning and how to design optimal testing instruments. Of necessity, this means that any teacher evaluation system must be phased in slowly to prevent penalizing teachers for things beyond their control.[401] States, LEAs, and schools must also be willing to play the long game and should not expect sudden improvements.[402] It is highly unlikely that student academic gains will be seen quickly, given that any new teacher evaluation system will be slowly implemented and gains will initially be limited.[403] As Eric Hanushek points out, even if every poor teacher were "deselected" immediately, it would take more than 10 years for student improvement to be fully realized—until a cohort of students completed all their schooling under the new system.[404]

D. *The Professionalization of Teachers*

As evaluation has increased, teacher autonomy has decreased. For example,

> In the 1920s, U.S. teachers' manuals emphasized the teaching of reading. The manuals were small, containing professional discussions . . . for the use of teachers. By the 1970s, teachers' lessons were completely scripted . . . to specify what to say, where to stand in

399. *See supra* Part III.
400. DOHERTY & JACOBS, STATE OF THE STATES 2015, *supra* note 328, at 2.
401. *See id.*
402. *See* Hanushek, *Teacher Deselection*, *supra* note 32, at 174.
403. *Id.*
404. *Id.*

classrooms, how to organize the lesson, and how to evaluate students.[405]

In addition, the very strategies that target improved learning increase teacher workload with the addition of or frequent changes to administrative tasks.[406] Whether each new "reform" offers greater possibility for student growth or whether it increases teacher workload to such an extent that it becomes undesirable has become an important consideration.[407]

As time has passed, responsibility for what (and how) to teach has become less the teacher's responsibility and more the product of administrative control.[408] Additionally, the growth of publication and testing firms that influence the content taught has taken some control out of teacher's hands.[409] While the government and special interest groups should have input into evaluation systems, wholesale evaluation transformation cannot be achieved without the input of those most affected by such changes—teachers themselves.[410] The lack of teacher input into evaluation systems cannot lead to lasting change; excluding teachers from important conversations sends the message that teachers either are not capable of or cannot be trusted to create fair and accurate evaluation systems. Additionally, a uniform national evaluation system is likely to be ineffective; instead, such systems should be developed by teachers and community leaders to tailor programs to the needs of the communities they serve.[411]

The history of teacher education has been one of marginalization; teachers have typically been granted little autonomy over their roles, as demonstrated by earlier historical models that designate teacher performance review as inspection and supervision rather than as evaluation.[412] Perhaps it is time that changed. Teachers have never been fully professionalized or made full participants in the educational process. Today's teachers must achieve professional status to be heard. Otherwise, the newest statutory changes will surely be decided without their input, and failure will result. Rather than feeling weakened by the changes to teacher

405. Popkewitz, *supra* note 75, at 4 (citation omitted).

406. *Id.*

407. *Id.*

408. *See id.*

409. *See id.*

410. NAT'L EDUC. ASS'N, TEACHER ASSESSMENT AND EVALUATION, *supra* note 366, at 3.

411. *Id.*

412. *See supra* Part II.

evaluations and increased accountability, teachers need to be more actively involved in designing evaluations that fairly and accurately measure their abilities.

E. *The Impact of Evaluation Models on Teacher Supply*

Any change to teacher evaluation—especially if connected to salary, tenure, and dismissal—is likely to have an impact on whether teachers are drawn to and remain in the profession. The current system rewards seniority, which is a benefit that keeps teachers in the system.[413] However, since teachers are not highly paid, the imposition of penalties or the elimination of current benefits may further limit the pool from which future teachers will be drawn. The use of bonus pay to improve teaching is equally problematic given that such bonuses must necessarily be small, thus disincentivizing teachers from earning them.

Nevertheless, as discussed in a previous section, penalty and reward structures are an important component of making accountability work, since "they have a positive impact on achievement."[414] It is unclear how best to reward teachers in a meaningful way so as to improve the overall quality of instruction. If rewards are restricted to high performers, then by necessity a small number of the overall teacher population will be rewarded and motivated. This must be balanced against the tension that penalties for lower performers may be difficult to impose and may discourage improvement. While eliminating the lowest, most persistently underachieving teachers from continuing in the profession is an important goal, it is undesirable to lose large numbers of teachers at once.

As teaching becomes more difficult and benefits decrease, it is possible that teachers will be driven into other professions. Any system overhaul will also likely be difficult to implement and sustain; the truth is that there are not enough top teachers to replace all middle and bottom performers,[415] no matter how good the idea.

F. *The Necessity for Improved Teacher Education and Training*

Given the many discussions and regulatory changes surrounding accountability, it may be time to consider the educational system that

413. U.S. Chamber of Commerce Found., *Teacher Compensation: Seniority Rules* (Mar. 2, 2011), https://www.uschamberfoundation.org/newsletter-article/teacher-compensation-seniority-rules.

414. Hanushek & Raymond, *supra* note 216, at 321; *see supra* Part III.

415. HAERTEL, *supra* note 30, at 7.

produces teachers, and whether it should raise barriers to entry to the profession in order to admit only those applicants who are most desirable. There are several ways to do this. One might be requiring higher test scores and grades from college applicants seeking to enter college or university teacher training programs. Another is requiring state or national board exams, including teacher observation and student achievement measures derived from student teaching. Any change to the current system must be balanced against the barrier it might create that would discourage excellent applicants from seeking a teaching career.

VAM is also just one method of improving overall teacher quality. Schools in the United States, compared to higher-performing international schools, have tenure and retention policies, which prevent the elimination of low-performing teachers.[416] In contrast, high-performing countries vary in their approaches to evaluation; some limit their teacher pool to high-performing college students while others require intensive teacher development.[417]

As stated above, change may also be needed in providing quality, focused teacher training post-hiring. This is an area of concern, as currently many teachers believe they are teaching with high academic quality. One study found that in the same time period in which the quality and quantity of teacher evaluations is increasing, less than half of 10,000 teachers participating in a 2015 survey believed that they needed improvement.[418] Since the greatest growth and increase in teacher quality tends to occur in the first five years of teaching, devoting resources to helping novice teachers improve could reap great benefits.[419] For teachers past the novice period, the quality of instruction tends to plateau and current teacher training and in-service programs do not have much effect.[420] Programs should be adopted that more effectively target areas of improvement for individual teachers,

416. SUSANNA LOEB ET AL., PERFORMANCE SCREENS FOR SCHOOL IMPROVEMENT: THE CASE OF TEACHER TENURE REFORM IN NEW YORK CITY 2 (2014), https://cepa.stanford.edu/sites/default/files/NYCTenure%20brief%20FINAL.pdf.

417. *See* ORG. FOR ECON. CO-OPERATION & DEV., TEACHERS FOR THE 21ST CENTURY: USING EVALUATION TO IMPROVE TEACHING 81–115 (2013), http://www.oecd.org/site/eduistp13/TS2013%20Background%20Report.pdf.

418. THE NEW TEACHER PROJECT, THE MIRAGE: CONFRONTING THE HARD TRUTH ABOUT OUR QUEST FOR TEACHER DEVELOPMENT 25 (2015), http://tntp.org/assets/documents/TNTP-Mirage_2015.pdf.

419. *See id.* at 14.

420. *Id.* at 2.

rather than provide generalized instruction that may not be applicable for all teachers attending school-provided training seminars.

G. *The Relationship Between Administrators and Teacher Evaluation*

Just as teachers are being held increasingly accountable for student achievement, so too must principals and administrators be held accountable for their ability to assess teacher performance accurately and provide feedback. VAM evaluations should not only affect individual teachers, but should also have consequences further up the administrative chain. Typically, the evaluation model for school principals is similar to that used for teachers in the same district; however, a component measuring the principal's ability to effectively assess teachers is typically omitted.[421] For example, 34 states require annual evaluations of school principals, but only 19 of those states weigh the ability to assess teachers accurately as main criteria of principal evaluation.[422] There is also a gap in training principals to perform teacher performance reviews correctly: slightly more than half of states require that principals be trained in teacher evaluation techniques to aid them in performing those important tasks.[423]

The number of evaluations per teacher per employment period has remained low. In 1999, 46 states relied on one evaluation conducted by either a principal or other designated person.[424] By 2015, only 11 states were requiring multiple evaluations of teacher quality.[425] This disconnect between the importance of evaluation and the necessity that quality evaluation be conducted is also highlighted when examining the evaluation policies surrounding principals' roles in the evaluation process—almost half the states fail to specify who should conduct such evaluations.

H. *The Need to Use Caution in Applying Business Employee Evaluation Models to Education*

Businesses have recently encouraged the remaking of teacher evaluation models to match their own practices, a reoccurring theme that is

421. DOHERTY & JACOBS, STATE OF THE STATES 2015, *supra* note 328, at iv–v.

422. *Id.* at iv.

423. *Id.* at iv–v.

424. David J. Wilkerson et al., *Validation of Student, Principal, and Self-Ratings in 360˚ Feedback® for Teacher Evaluation*, 14 J. PERSONNEL EVALUATION EDUC. 179, 181 (2000).

425. DOHERTY & JACOBS, STATE OF THE STATES 2015, *supra* note 328, at 14.

more than 100 years old.[426] At several points in the history of teacher supervision and evaluation development, business models were proposed as ways to measure teachers to determine whether they were as productive as a factory worker manufacturing a set number of parts per hour.[427]

Part of the pressure to adopt VAM as an evaluation model for teachers comes from the perception that businesses and professions evaluate their employees in this way for accuracy and efficiency purposes.[428] However, there is no true corollary that can be drawn between these disparate jobs.[429] The quantitative measures used for other job evaluations tend to have more to do with total sales, sales volume, or success rates.[430] Rarely do these measures actually rely on data generated by testing the customer. Furthermore, there is no other job that attempts to require workers to show the level of love and understanding to customers that teachers are encouraged to show to students. Therefore, defining a student as a "product" of education is an oversimplification at best.

VII. CONCLUSION

Determining the most appropriate teacher evaluation model requires us to determine our educational goals and to remember that better test performance does not equal better teaching and learning. This Article has reviewed the history of U.S. teacher evaluation models,[431] looked at evaluation in the modern era,[432] and identified policy considerations for future models.[433]

While VAM has been widely adopted, it must be tied to other teacher personnel policies to be effective.[434] Teachers must also receive evaluation feedback in a timely way so that they can respond quickly.[435] Of necessity, any potential teacher evaluation model that relies on student achievement

426. *See* Tracy, *supra* note 59, at 323 (discussing the impacts of early twentieth century industrialization on education).

427. *See, e.g.*, *id.*

428. BAKER ET AL., *supra* note 38, at 6.

429. *See id.*

430. *See id.* at 6–7.

431. *See supra* Parts II–III.

432. *See supra* Parts IV–V.

433. *See supra* Part VI.

434. Hanushek & Raymond, *supra* note 216.

435. *See supra* Part IV.

indicators must rely on accurate data.[436] Given the variability that can occur, it is likely that one or two years of data is not enough to indicate teacher success or failure. Nonetheless, in states where tenure decisions are made rapidly, one year of student growth indicators is likely all that administrators will see before extending an offer for a permanent position. Thus, the only way to evaluate successfully is to determine whether the goal of the evaluation model is improving teaching quality or assessing the need for teacher dismissal.

In adopting VAM, attention should continue to be paid to disadvantaged students. While a core belief in the United States is that education is the great equalizer, change will be required to make that true.[437] Students born into poverty perpetuate the cycle, as do students born into wealth. To support the belief that education creates great opportunity, we need to make sure school quality is high for all students. We also need to discover how best to help students at the bottom who may need more than instruction to help them succeed.

Finally, school accountability and teacher accountability cannot neatly be divorced from one another as each is dependent on the other for success.[438] The school must bear some responsibility for hiring, training, and allocating resources to individual classrooms. It is clear that changes in the teacher evaluation process are necessary and underway. What is less clear is whether the introduction of new evaluation systems will bring about the reforms sought by advocates. There is no one answer to solve our educational problems. Teacher evaluation and its proposed concomitant uses are only steps toward eventual educational success.

436. *See supra* Part IV.
437. *See supra* Part VI.B.
438. *See* discussion *supra* Part VI.C.