The Report Committee for Nicole Joanne Moore
Certifies that this is the approved version of the following report:

Investigating the Use of Value-Added Models for Student Achievement: Does Using Multiple Value-Added Measures Lead to Stronger Conclusions about Teacher Effectiveness?

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Cynthia Osborne

Paul von Hippel

Investigating the Use of Value-Added Models for Student Achievement: Does Using Multiple Value-Added Measures Lead to Stronger Conclusions about Teacher Effectiveness?

by

Nicole Joanne Moore, BBA

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Public Affairs

The University of Texas at Austin

May 2012

Dedication

I dedicate this report to the countless teachers who believed in me and inspired me to become who I am today. From Ms. Blessington in Kindergarten who helped me learn to read to Ms. Ballard in high school who encouraged my post-secondary pursuits, my teachers taught me the power of believing in myself. May we always support teachers to inspire greatness in their students and bring greater equity to our world.

Acknowledgements

My undying gratitude goes to my amazing friends and advisors who supported my work and didn't let me give up on this project. Talitha offered invaluable support in providing a framework for working on this massive undertaking and offering encouragement along the way. Dr. von Hippel and Dr. Lincove provided great insight about the right kinds of questions to be asking in the quest for better educational outcomes. Dr. Osborne continually pushed me to a higher level of excellence than I've known before, and for that she has my gratitude. Working with the Project on Educator Effectiveness and Quality opened my eyes to the complex statistical processes and policies around value-added measures and enabled this research in many ways. Many thanks to the large urban school district that provided my K-12 education, inspired this topic, and also offered data and support along the way. Jonathan Sims provided encouragement in the most trying parts of this process and will forever hold a special place in my heart. I would never be where I am today without the support of all those who love me, believe in me, and inspire me daily to make an impact in the world.

Also, a special shout out goes to Florence and the Machine, Ray LaMontague, Peter Bradley Adams, Ingrid Michaelson, Of Monsters and Men, and all the other musicians whose playlists inspired my work.

Abstract

Investigating the Use of Value-Added Models for Student Achievement: Does Using Multiple Value-Added Measures Lead to Stronger Conclusions about Teacher Effectiveness?

Nicole Joanne Moore, MPAff

The University of Texas at Austin, 2012

Supervisor: Cynthia Osborne

In the quest to achieve better academic outcomes for all students, the focus in education has shifted to a model of accountability. The most recent trend in the accountability movement is a focus on the effect of teachers in promoting student achievement. Research has found that teachers have the most significant school level impact on student achievement, and increases in teacher effectiveness could have major implications for the learning outcomes of students across the nation. Much of the current focus in teacher evaluation reform centers on methods through which teachers can be more accurately evaluated based on their contributions to student learning. In the push towards greater accountability for teachers, the development of measures that are both fair for teachers and lead to stronger outcomes for students are critical to seeing long-term improvements in the education system.

This report explores variability and stability of value-added measures over time by looking in depth at the methods, assumptions, limitations, and implementation of the most commonly used value-added models across the country and the research about the correlations of these measures over time. This research is followed by a case study of a de-identified large urban school district implementing a teacher evaluation system that uses both a commercially produced value-added measure and an alternative student-growth measure to make high stakes decisions about teacher effectiveness. The findings from this case study show correlations that do not differ significantly from the prior research on the year-to-year variability in teacher value-added measures, but urge for continued evaluation of these measures over time, especially in high-stakes decisions. Ultimately, value-added measures are only as useful as their effectiveness in influencing the core outcomes of teaching and learning, and therefore these measures must be carefully integrated into and validated against holistic assessments of teacher effectiveness in order to truly impact student outcomes.

Table of Contents

List of Tables

Introduction

The evolving agenda of education reform centers on a fundamental question: what can be done to improve student outcomes? While most stakeholders agree that all students deserve a chance to succeed, this core question seeks a solution to fix a system that fails to bring educational opportunity to all students. On the National Assessment for Education Progress (NAEP) test in 2009, 67 percent of 4th grade students and 68 percent of 12th grade students earned below proficient scores in reading. This trend of low academic achievement also appears in mathematics, where 61 percent of 4th grade students and 74 percent of 12th grade students earned below proficient scores. These statistics are even more striking when disaggregated by racial group. On the NAEP 2009, black 4th grade students earned average reading scores 26 points below their white peers. In the 12th grade mathematics assessment in 2009, white students outperformed black students by 30 points and Hispanic students by 23 points.[1] Trends in education reform emerge as an answer to this fundamental question: what can be done to substantially impact the education system to get better results?

One of the recent responses to this question is the rise of accountability measures designed to hold schools and teachers responsible for the educational outcomes of their students. The No Child Left Behind Act (NCLB) required schools to meet proficiency standards to maintain funding and recognition. The NCLB system was intended to measure student and school achievement in a systematic way, with a goal that all students would be able to reach proficiency. With a greater awareness of student outcomes, NCLB

[1] National Center for Education Statistics. The Condition of Education. Retrieved on April 25, 2012 from http://nces.ed.gov/programs/coe/indicator_mgp.asp.

proponents hoped that schools would be better able to provide support for struggling students and would be held accountable for reaching academic goals.

This move towards test-based accountability represents a significant shift in education trends, as it provides a mean to systematically evaluate the core goal of student outcomes.[2] This emphasis on school-level accountability has more recently shifted to the teacher-level, holding individual teachers accountable for the academic achievement of their students. Instead of focusing on a school, the focus on teachers brings accountability to an individual level. Through factors outside of school explain the largest percent of the variance in a student's education outcomes, teachers have the largest effect on student outcomes at the school level. Research has found wide amounts of variation in effectiveness among educators even at a single school.[3] However, as with any accountability measure, the underlying assumptions and implications for teachers must be carefully examined in order for the measures to have the desired impact on students and schools.

Despite the desire for expedient results in education reform, newly-enacted policies must still be carefully evaluated for possible unintended consequences that may actually inhibit the core goals of teaching and learning. This report adds to the current research by specifically examining the stability of value-added models used to measure teacher effectiveness over time and the consistency of these measures between subjects and years. Recent education reforms have embraced value-added models as means to estimate the unique contributions of the school or teacher on students' progress over the

---

[2] National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[3] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 418.

course of a year rather than the cumulative effects of education or student background factors. Understanding the possible implications from the use of these measures in high stakes evaluation decisions is vital to be ensure that value-added measures are used for the ultimate goal of education: to improve student outcomes. Ultimately, value-added models may not lead to stronger conclusions about teacher effectiveness due to inconsistency in these measures over time; therefore, the use of value-added models for teacher evaluation should be validated and complemented by more comprehensive measures of teacher effectiveness.

To examine the use and implications of value-added measures, the report is organized into five chapters. Chapter 1 discusses the trends in the accountability movement since the authorization of the Elementary and Secondary Education Act, with a specific focus on increased accountability for teachers in promoting student achievement. This chapter also examines the research on a teacher's impact on student achievement and discusses the implications of focusing on teacher effectiveness for student learning outcomes. Chapter 1 also explores the development of new measures for evaluating teacher effectiveness, which are more closely tied to a teacher's contributions to student learning.

Chapter 2 explores the use of educational value-added measures, which seek to quantify the unique contribution of teachers to student achievement, based on student test scores. This chapter begins with a theoretical framework for value-added measures and then explores the key empirical and policy-level challenges with these measures. Chapter 2 continues with an examination of the methods, assumptions, limitations, and implementation of four of the most commonly used models across the country, with a specific focus on the purpose and reliability of the measures.

Chapter 3 highlights a summary of the research findings about the correlation and consistency between various value-added models. This chapter examines four studies that compare value-added models between subjects and over time and examines the key research questions, findings, and implications from this research.

Chapter 4 includes a case study of a de-identified large urban school district implementing a teacher evaluation system that uses both the Education Value-Added Assessment System and an alternative student-growth measure to make high stakes decisions about teacher effectiveness. This chapter discusses the background of the district's work with these value-added measures and then describes the research methodology and findings from the correlation analysis of the district's data.

Chapter 5 concludes with recommendations and the implications for using multiple models and how value-added models can most effectively be used to improve schools and student outcomes. This chapter explores the implications of the ideal use of value-added measures and possible consequences for teachers and school districts if the measures aren't carefully used. Ultimately, value-added measures are only as useful as their effectiveness in influencing the core outcomes of teaching and learning, and therefore these measures must be carefully integrated into and validated against holistic assessments of teacher effectiveness in order to truly impact student outcomes.

Chapter 1: Policy Background

ACCOUNTABILITY FOR RESULTS IN SCHOOLS

Test-based accountability has been one of the most enduring policy reforms in the field of education. The focus on accountability in education has been in place since the 1800s, when the first standardized testing was used as a more objective basis for measuring student knowledge. The 1920s represented the peak of the "scientific management" movement, which attempted to improve efficiency in all types of organizations through psychological testing of knowledge and thinking skills.[4] Testing for accountability purposes continued under Title I of the Elementary and Secondary Education Act (ESEA) of 1965 and with the creation of the National Assessment of Educational Progress (NAEP) in 1969.[5] The original form of these national testing requirements did not include explicit incentives or accountability linked to test results.

Since the release of Coleman's report of Equality and Education Opportunity in 1966, the education policy debate in the United States has included a discussion of the role of schools in producing student achievement.[6] This report found very small effects of differences in the measured attributes of schools on student achievement when compared to a student's background factors, such as socio-economic status. The Coleman Report and similar research calls into question the extent of the relationship between education-related factors and learning outcomes.[7] Some research has suggested that "schools bring

---

[4] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 24.

[5] National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[6] Coleman, J. et al. (1966). Equality of Educational Opportunity. National Center for Education Statistics. Retrieved from http://www.eric.ed.gov/PDFS/ED012275.pdf.

[7] Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy*

little influence to bear upon a child's achievement that is independent of his background and general social context."[8] Within these frameworks, the relative impact of teachers and schools may be severely limited in comparison to other factors.[9]

Despite questions about the relative significance of schools in producing student achievement, the drive towards school accountability grew in following decades, fueled in part by the publication of *A Nation at Risk* in 1980. The publication of this report represented a shift in the national approach towards education and emphasized the importance of content standards, which formed the basis for the expansion of testing. The 1988 reauthorization of ESEA required Title I, high-poverty schools with stagnant or declining test scores to file improvement plans with their districts. In the 1990s, the federal government continued the shift towards "new accountability" by requiring school report cards with average student test scores.[10]

The standards-based reform movement of the early 1990s led to the requirement in the 1994 ESEA reauthorization for states to create rigorous content and performance standards and report student test results in terms of the standards. The growing interest in tying student learning to educational accountability has stimulated unprecedented efforts to use high-stakes tests in the evaluation of individual teachers and schools.[11] Test-based accountability has taken even greater hold of education policy in the first decade of the 21st century through the enactment of the No Child Left Behind Act (NCLB), the state

---

Analysis Archives, *8*(1). Retrieved from http://epaa.asu.edu/ojs/article/view/392, 2.

[8] Ibid.

[9] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 418.

[10] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 24.

[11] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.

movement for high school exit exams, and the development of value-added measures to tie teacher pay to student test results.[12]

The No Child Left Behind Act, signed into law by President Bush in 2002, represented a significant shift in federal education policy as the federal government became a major force in shaping the goals and outcomes of education. The legislation was fueled in part by the seeming ineffectiveness of Title I federal expenditures, which gave funding to schools based solely on the number of students regardless of student performance. NCLB established a comprehensive framework of standards, testing, and accountability and removed some discretion from local education authorities in determining what the goals and outcomes of education should be.[13] The initial framework required the yearly testing of all students in grades 3 through 8 in reading and math and set mandatory adequate yearly progress goals for all schools. This law set a clear emphasis on results, with the ultimate goal that all students would achieve proficiency by 2014. NCLB also included requirements about the reporting of results by student subgroups, broken down by ethnicity, special education, English-language learners, and economic disadvantage. The promise of NCLB to enhance equity and opportunity by reducing the achievement gap fell short due to both insufficient funding and an overly simplistic definition of the achievement gap.[14]

NCLB also adopted a very narrow definition of teacher quality, which has resulted in a tension between decision makers and professional educators over what

---

[12] National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
[13] Fusarelli, L. (2004). The Potential Impact of the No Child Left Behind Act on Equity and Diversity in American Education. *Education Policy, 18:71*. Retrieved from http://sitemaker.umich.edu/tabbye.chavous/files/fusarelli2004.pdf.
[14] Ibid.

constitutes an excellent teacher.[15] In this act, teacher quality was formalized as a set of minimum qualifications that teachers must achieve before becoming eligible to teach.[16] The law set requirements that all teachers of core academic subjects be "highly qualified," including the minimum requirements of a bachelor's degree, full state licensure and certification, and demonstrated subject-area competence.[17] However under the NCLB definition, teacher quality did not include measures of a teacher's impact on student achievement and didn't differentiate the impact of a teacher versus a school on student outcomes.

Differentiating between School and Teacher Accountability

The focus on school accountability has more recently shifted to the teacher level, with a focus on measuring the impact that individual teachers have on student achievement. It may be difficult to measure and differentiate between school and teacher contributions to student learning. A multitude of school-level factors affect educational outcomes, including class sizes, staff support, school and district leadership, funding for textbooks and supplies, and community support, which are largely outside teachers' control.[18] Research has found that the most important variables at the school level are the staff's value-orientations: teachers' belief in their students' ability to learn, high expectations, and discriminating reinforcement of learning behavior.[19] School

---

[15] Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge, 1.
[16] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf
[17] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution.
[18] Ibid, 5.
[19] Campbell, J., Kyriakides, L., Muijis, D., and W. Robinson. (2004). *Assessing Teacher Effectiveness: Developing a differentiated model*. RoutledgeFalmer, 7.

effectiveness may also include factors such as leadership, school climate, and school policies that contribute to student performance. Accounting for influential school-level factors is important in disentangling school and teacher effects on student achievement.

Teacher effectiveness may be distinguished from school effectiveness as the impact that classroom factors, such as teaching methods, teacher expectations, classroom organization, and use of classroom resources have on student performance. The concept of teacher effectiveness may be difficult to singularly define, but some researchers have proposed it to be "the power to realize socially valued objectives agreed for teachers' work, especially the work concerned with enabling students to learn."[20] One survey in 2009 found that 76 percent of teachers believed that making it easier to dismiss ineffective teachers would improve teacher effectiveness, and 32 percent believed that tying rewards such as salary to measured performance would do the same.[21] These findings suggest that the development of effective measures and consequences in assessing teacher effectiveness is essential to reaching desired outcomes in improving student achievement.

Although accountability is an important goal in education, attempts to focus only on the teacher may ignore critical, interrelated parts of the educational process.[22] A definition of teacher effectiveness as simply a teacher's ability to improve student learning as measured by student gains on standardized achievement tests seems a narrow

---

[20] Ibid, 4.

[21] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 5.

[22] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 15.

way to assess the impact of teachers.[23] Student learning is impacted by a variety a different factors beyond a single teacher including other teachers, peers, family, home environment, poverty, school resources, community support, leadership, and school climate. Growth in student social development such as improvement in student attitudes, motivation, and confidence also contributes to learning in ways that may not appear in results on standardized tests.[24]

FOCUS ON TEACHER EFFECTIVENESS

As the closest point of school-level influence on students, the over three million teachers in elementary and secondary schools across the country play a huge role in the success of the U.S. public education system.[25] The recent focus on measuring teacher effectiveness comes in light of research that shows the significant impact of teachers on student achievement outcomes. Research studies have found that teachers have the largest influence on student learning at the school level and that variation in teacher quality has a differential impact on student achievement scores.[26] A study from Los Angeles Unified School District found that the average student assigned to a teacher who was in the bottom quartile of performance during his or her first two years lost on average five percentile points relative to students with similar baseline scores and demographics. The reverse impact was also found for students who were assigned to a top quartile teacher, who on average gained five percentile points relative to students with similar baseline

---

[23] Little, O., Goe, L., & Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness.* Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/practicalGuide.pdf.

[24] Ibid.

[25] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution.

[26] Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *National Bureau of Economic Research Working Paper Series, No. 14607.*

scores and demographics. These research findings suggest that having a top-quartile teacher rather than a bottom-quartile teacher four years in a row with persistent and accumulated effects would be enough to close the black-white test score gap.[27] The research on teacher effectiveness also finds that the quality of teachers has been found to vary across schools in a way that systematically disadvantages poor, low-achieving, and racially isolated schools.[28] The distribution and effectiveness of teachers has influential effects on the quality of education that students receive.

Numerous other research studies also have substantiated the differential effects to student outcomes from a highly effective teacher. Researchers using data from Chicago public high schools found that having an instructor who was rated in the 95th percentile in teacher quality could add 25 to 45 percent of an average school year's growth to a student's mathematics score.[29] A study using data from the Tennessee Project STAR conducted an analysis of teacher effects, defined as the portion of student achievement that remains unaccounted for after controlling for student demographics, class size, and school level effects. The researchers found significant teacher effects on achievement gains for both the mathematics and reading tests.[30] Other recent studies of teacher effects at the classroom level have found that differential teacher effectiveness is a strong

---

[27] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution, 8.
[28] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf
[29] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 40.
[30] Ibid, 41.

determinant of differences in student learning, far outweighing the effects of differences in class size and variance among student achievement outcomes.[31]

Although there is much research supporting the impact of effective teachers on student achievement, less research has explored the decay of teacher effects over time. One study that involved a random-assignment experiment in the Los Angeles Unified School District found that teacher effects fade out by roughly 50 percent per year in the two years following teacher assignment.[32] While a student may achieve significant gains under a single effective teacher, these gains decay substantially over time. Other researchers, such as Jesse Rothstein, have questioned the causal relationship between long-term student achievement and teacher value-added scores. Rothstein's falsification test has shown that in some cases student scores that regress towards the mean may lead to overstated teacher effects the subsequent year.[33] Additional research findings on the persistence of value-added teacher effects over time are explored more fully in chapter 3 of this report, and these findings underscore the need for balanced estimation of how increases in teacher effectiveness can influence long-term student achievement.

A recently published report that followed the long-term impact of teachers found significant correlation between students of teachers with high value-added scores and stronger life outcomes. This research analyzed school district and tax records for 2.5 million children in grades 3-8 and found that students assigned to high value-added teachers are slightly more likely to attend college, attend higher-ranked colleges, earn

---

[31] Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *8*(1). Retrieved from http://epaa.asu.edu/ojs/article/view/392, 2.

[32] Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *National Bureau of Economic Research Working Paper Series*, *No. 14607*.

[33] Rothstein, J. (2008). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w14442.pdf.

high salaries, live in higher SES neighborhoods, and save more for retirement. They are also less likely to have children as teenagers. The study found that on average, one standard deviation improvement in teacher value-added in a single grade is associated with a 1% increase in earnings at age 28.[34] The emerging research about the differential impact of teachers on student achievement highlights the potential economic value that effective teachers create and underscores the opportunity to develop more informed evaluations of teacher quality and accountability.

Defining Teacher Quality

While there is general agreement that teacher quality matters in terms of student achievement, no clear consensus exists on which aspects of teacher quality matter most in forging a useful definition of teacher quality.[35] Despite the current research on the impact of teachers in student achievement, until recently many of the traditional methods for teacher evaluation didn't include student performance measures as a significant component. The New Teacher Project produced a report entitled "The Widget Effect" in 2009, which detailed the lack of differentiation in teacher evaluation results. The study found that within districts that use binary evaluation ratings (typically "satisfactory" or "unsatisfactory"), more than 99 percent of teachers receive the satisfactory rating. The report concludes that in many districts, a teacher's effectiveness "is not measured, recorded, or used to inform decision-making in any meaningful way."[36] Without

---

[34] Chetty, R., Friedman, J. & J. Rockoff. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. National Bureau of Economic Research. Retrieved from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf.

[35] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 1.

[36] Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. *The New Teacher Project*. Retrieved from http://widgeteffect.org/

informed evaluation practices for teachers, these yearly assessments of teacher practice serve a primarily perfunctory role without any impact on teaching or learning outcomes.

Besides rating almost all teachers as satisfactory, most current models for teacher evaluation fail to provide any useful guidance or support to help teachers improve their practice. Randi Weingarten, president of the American Federation of Teachers, noted in a recent speech "Our system of evaluating teachers has never been adequate. For too long and too often, teacher evaluation – in both design and implementation – has failed to achieve what must be our goal: continuously improving and informing teaching so as to better educate all students."[37] This credentialing strategy of teacher effectiveness – a focus on tenure, single-salary schedule, checklist evaluations, and certification – fails to address the core purpose of teaching: improving student outcomes.[38]

Another common method used to differentiate teacher quality is through various input measures, such as advanced degrees, years of teaching experience, and certification. Research has found that these input factors may have only a weak impact on teacher effectiveness, and in some cases are negatively correlated.[39] The studies on teaching experience have suggested that increases in effectiveness level beyond the fifth year of teaching contribute little or no additional benefit in terms of student achievement.[40] One study using data from Los Angeles found that the return to the first few years of experience is less than half as large as the difference between the highest and lowest

---

[37] Weingarten, R. (January 2010). *A New Path Forward: Four Approaches to Quality Teaching and Better Schools*. Speech presented at Education Quality for the 21st Century. Washington, D.C.

[38] Ibid.

[39] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution.

[40] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 3.

performing quartiles of teachers in their first two years.[41] Differences in effectiveness between certified and uncertified teachers also do not have a significant impact on student outcomes. In a study from Los Angeles Unified School District, researchers found the difference between the a teacher at the 50th percentile and 75th percentile among all teachers was roughly five times as large as the difference between the average certified teacher and the average uncertified teacher. This gap was roughly the same as the difference between the 25th percentile teacher and the 50th percentile teachers, which shows there is wide variation across teacher effectiveness beyond certification status.[42] To compile a more complete picture of teacher effectiveness, there needs to be a differentiation between teacher quality, the set of inputs that indicate a highly qualified teacher, and teaching quality, which is based on what results teachers get in the classroom.[43] Using input measures alone as a means to define teacher quality produces an incomplete reflection of teacher effectiveness.

TEACHER EVALUATION REFORM

As a result of the findings on the impact of teachers on student performance, many states and districts have shifted to a new paradigm of measuring effectiveness on the basis of student outcomes as opposed to teacher inputs.[44] Leaders of both political parties have endorsed linking teacher evaluation to student test scores, a dramatic shift

---

[41] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution, 28.
[42] Ibid, 7.
[43] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 8.
[44] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf

from previous evaluation policies. In promoting the national Race to the Top program, President Obama stated, "Success should be measured by results, and data is a powerful tool to determine results… That's why any state that makes it unlawful to link student progress to teacher evaluation will have to change its ways. The Race to the Top grants will go to states that use data effectively to reward effective teachers, to support teachers who are struggling, and when necessary, to replace teachers who aren't up to the job."[45] The national focus on improving teacher quality has fueled innovation and reform through multiple public and private initiatives, which are designed to more accurately measure and incentivize teacher effectiveness.

One of the influential factors in these reforms is Race to the Top, an initiative of the Department of Education that made nearly $4.4 billion available to fund education reform at the state level. This program was announced in July 2009, and eleven states and Washington, D.C. won grants in the first two rounds. The Race to the Top program focuses on four specific areas: adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy; building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction; recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and turning around our lowest-achieving schools.

One of the Race to the Top priorities is the incorporation of student performance as a significant factor in teacher evaluations and in decisions regarding hiring, firing, tenure, and compensation.[46] Race to the Top defines "highly effective teachers" as those

---

[45] Obama, B. (July 2009). *Remarks by the President on Education*. Speech presented at the Department of Education. Washington, D.C.

[46] Buckley, K. & S. Marion. (2011). A Survey of Approaches Used to Evaluate Educators in Non-tested Grades and Subjects. National Center for the Improvement of Educational Assessment.

who students achieved high rates of growth, defined by the program as a change in test scores between two or more points in time.[47] In response to the Race to the Top program, many states removed data "firewalls" that have prohibited educators from linking student achievement to individual teachers.[48] Many of the grant winners are incorporating value-added models of student growth as a means to evaluate teacher effectiveness.[49] One of the larger goals of the program is to fuel nationwide education reform through the replication of successful initiatives from winning states. A number of the provisions from Race to the Top may be integrated as part of the reauthorization of the federal Elementary and Secondary Education Act.

Another national initiative designed to improve teacher quality is the teacher incentive fund (TIF), which supports efforts to develop and implement performance-based teacher and principal compensation systems in high-need schools. The goals of the program include improving student achievement through increasing teacher and principal effectiveness and creating sustainable performance-based compensation systems.[50] The Department of Education has committed $1.2 billion over the next five years to this fund. These grants were awarded in part by plans to create and implement several measures to identify and reward effective teachers using measures of student growth.[51] Several TIF

Retrieved from http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf, 2.

[47] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf

[48] Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge, 15.

[49] Department of Education. Race to the Top Fund. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

[50] Department of Education. Teacher Incentive Fund. Retrieved from http://www2.ed.gov/programs/teacherincentive/index.html.

[51] Buckley, K. & S. Marion. (2011). A Survey of Approaches Used to Evaluate Educators in Non-tested Grades and Subjects. National Center for the Improvement of Educational Assessment.

awards have gone to districts implementing the Milken Foundation's Teacher Advancement Program (TAP), which provides multiple career pathways within schools, ongoing professional development, instructionally focused accountability, and performance-based compensation. Approximately 6,000 teachers in 50 school districts nation-wide participate in this program. TAP uses a value-added model to determine contributions to student achievement gains at both the classroom and school levels. Teachers are awarded bonuses based on an evaluation that includes mastery of effective classroom practices, student achievement gains, and school-wide achievement gains.[52]

One of the largest privately funded responses to improving teacher effectiveness is the Measures of Effective Teaching (MET) Project from the Bill and Melinda Gates Foundation. Launched in fall of 2009, The MET Project is based on three premises: first, a teacher's evaluation should include his or her students' academic achievement gains; secondly, any additional components of the evaluation should be valid predictors of student achievement gains; and thirdly, any measure should include feedback on specific aspects of a teacher's practice to support teacher growth and development.[53] To identify the most significant measures of teacher effectiveness, the project is working with over 3,000 teachers in six predominately urban districts across the country. The project is collecting data on student achievement gains on various assessments, classroom observations and teacher reflections, measures of teachers' pedagogical content

---

Retrieved from http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf, 2.

[52] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 20.

[53] Measures of Effective Teaching Project. (2010). *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.gatesfoundation.org/highschools/Documents/met-framing-paper.pdf

knowledge, student perceptions of the classroom instructional environment, and teachers' perceptions of working conditions and instructional support at their schools.

The initial findings from the multi-year study have found that a teacher's past track record of value-added scores is among the strongest predictors of their current students' achievement gains. The study also found that student perceptions in one class are related to the achievement gains in other classes taught by the same teacher. The final goal of the project is to improve the quality of information about teacher effectiveness in order to help build fair and reliable systems for teacher observation and feedback.[54]

National Responses

The push of innovation in calculating teacher effectiveness was also recently spotlighted in national headlines out of Los Angeles. Using a freedom of information request, the L.A. Times gathered seven years of reading and math scores from the L.A. Unified School District and calculated the performance of over 6,000 teachers who had taught Grades 3 through 5. These results excited a national fervor over calculating and releasing individual-level value-added results for teachers.[55] While this type of analysis was not novel, the real controversy came from the newspapers publishing of the individual teachers' value-added scores along with their names. This release of teacher rankings was replicated in New York City, where the ratings of 18,000 teachers were published in February 2012. These teacher data reports covered three school years and were intended to show how much value individual teachers add by measuring how much

---

[54] Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

[55] Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge, 14.

their students' test scores exceeded or feel short of expectations based on demographics and prior performance. Although the city's Education Department stated that these value-added measures were not intended to be used in isolation, the results were published on basis of value-added ranking alone.[56] The reporting of these results contributes to educator mistrust of policy makers to design appropriate accountability policies and of the media to accurately portray school performance.[57] Although states and districts across the nation are working to improve teacher effectiveness and student educational outcomes, the methods of calculating and sharing this information are vital in ensuring their success.

MEASURING STUDENT ACHIEVEMENT

Under No Child Left Behind, school accountability measures required yearly increases in the percentage of proficient students to assess if adequate yearly progress had been made. This method of evaluating student performance led to the development of assessments that were designed to measure a minimum standard of student knowledge and proficiency. These measures also did not factor in possible changing student demographics or how to account for schools that regularly showed high levels of student achievement. The NCLB accountability guidelines resulted in many unintended consequences that did not necessarily contribute to increased student outcomes. In fall of 2011, the U.S. Department of Education offered NCLB waivers for states in exchange for rigorous and comprehensive state-developed plans designed to improve educational

---

[56] Santos, F. & R. Gebeloff. (2012, February 24). Teacher Quality Widely Diffused, Ratings Indicate. *New York Times*. Retrieved on March 27, 2012, from http://www.nytimes.com/2012/02/25/education/teacher-quality-widely-diffused-nyc-ratings-indicate.html?_r=1&ref=robertgebeloff.

[57] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 3.

outcomes for all students, close achievement gaps, increase equity, and improve the quality of instruction.[58]

In the push for school accountability, the most commonly used methods to measure student achievement are status models, cohort-to-cohort change models, growth models, and value-added models. Each type of model is designed to answer a set of policy-relevant questions.[59] Status models show a snapshot of student performance at a point in time, which can be compared with an established target. A status model is the traditional measure used under No Child Left Behind and answers the question "Has school X met the state proficiency target this year?" Cohort-to-cohort change models measure the change in test results for a teacher, school, or state by comparing status at two points in time, although not for the same groups of students. Under NCLB, this measure is commonly used to answer the question "Are students at a certain grade level doing better this year in comparison to the students who were in the same grade last year?" Growth models measure student achievement by tracking the test scores of the same students from one year to next to determine the extent of their progress. This model answers the question "How much, on average, did students' performance change between grade X and grade Y?" Many accountability systems may set a target for an expected amount of growth for schools or subgroups of student. The fundamental question in choosing a method to measure student achievement is how a state or school district defines success. Based on the underlying goals of the school system, a state or district may choose an aligned method for measuring student achievement and growth.

---

[58] Department of Education. ESEA Flexibility. Retrieved from http://www.ed.gov/esea/flexibility.
[59] Braun, H. Chudowsky, N., and Koenig, J. eds. (2010). Getting Value Out of Value-Added: Report of a Workshop. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability; *National Research Council*. Washington, D.C.: The National Academies Press, 4.

Although many states and districts rely on status models, cohort-to-cohort change models, or growth models for measuring student achievement, a few are exploring the use of more complex models that use longitudinal data on students to determine the "value added" by a particular teacher or school.[60] Value-added results refer to "efforts to estimate the relative contributions of specific teachers, schools, or programs to student test performance."[61] Unlike a set proficiency bar, these methods seek to isolate the portion of a student's success that cannot by attributed to any other current or past student, school, family, or community influence.[62] Controlling for at least student prior test scores, value-added models calculate an expected score for a student so that the difference between the actual gain score and the predicted gain score can be positively or negatively attributed to the teacher. These newly developed models are quickly becoming the leading approach for holding teachers accountable for student performance on standardized assessment results.[63]

Value-added models have stood at the centerpiece of a national movement to evaluate, promote, compensate, and dismiss teachers based in part on their students' test results. Support for the value-added approach in education accountability has stemmed in part from the belief that it can remove the effects of factors not under the control of the

---

[60] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 68.

[61] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press. *vii*.

[62] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf

[63] Buckley, K. & S. Marion. (2011). A Survey of Approaches Used to Evaluate Educators in Non-tested Grades and Subjects. National Center for the Improvement of Educational Assessment. Retrieved from http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf, 3.

school, such as prior performance and socioeconomic status, and thereby provide a more accurate indicator of school or teacher effectiveness than is possible when these factors are not controlled.[64] Federal, state, and local policy-makers have been drawn to these measures in an attempt to objectively quantify teaching effectiveness and promote and retain teachers with a demonstrated record of success.[65] The following chapter will give a more in-depth look at different types of value-added models used across the country, and the benefits and challenges of these methods when used in teacher evaluation.

[64] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 68.
[65] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 4.

Chapter 2: A Comparison of Value-Added Models

Overview of Value-Added Models

Value-added models in education are used to estimate the unique contributions of the school or teacher on students' progress over the course of a year rather than the cumulative effects of education or student background factors.[66] The appeal of value-added models is that these measures calculate student growth based on a student's achievement pattern over time, rather than measuring student proficiency based on an absolute bar. This focus on student achievement gains rather than differences in test scores allows each for student's prior testing history to be controlled for in the model. The isolation of the effects of educational and other factors is critical for drawing accurate conclusions about teacher effectiveness and may be key to making significant improvements in education.[67]

Value-added modeling uses statistical methods to analyze students' prior test scores and make predictions for student performance over time. To approximate a value-added result, researchers need test data from at least two points in time for each student in the same subject to measure predicted and actual student gains. As part of the requirements of No Child Left Behind, all states administer tests in 3-8 in English Language Arts and Mathematics. Although these state accountability tests were designed to measure student proficiency, many states and districts are using these assessments to calculate teacher contributions to student learning. The reliance on these state

[66] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 1.
[67] Ibid, 2.

assessments to find teacher value-added also means that the 65-75% of teachers who do not administer a standardized test require other methods for evaluation.[68]

Limitations of Value-Added Models

Although value-added models can give a more nuanced picture of student growth, these measures also have many limitations in both their empirical basis and policy implications, including issues with error and bias, the use of standardized tests, the choice of variables in the model, and concerns a lack of transparency with complex statistical models. One of the primary issues in using value-added models is whether or not they are able to truly isolate a teacher's unique effects on student learning. A value-added model must be carefully specified to account for other factors that influence student achievement and provide an estimate of the unique teacher effect. Also, in order to be an accurate measure of teacher effectiveness, researchers need a high level of confidence in the attribution of achievement gains to specific teachers. In most value-added models, each teacher has a confidence interval representing the level of certainty associated with the value-added percentile measure, which accounts for the possible error in the model.[69]

Other confounding factors in value-added modeling include the assumption that student's test performance is equated with their knowledge of the subject, even though their performance may be affected by other influences such as motivation, test-taking strategies, and attitudes toward testing. In addition, value-added models average the

---

[68] Buckley, K. & S. Marion. (2011). A Survey of Approaches Used to Evaluate Educators in Non-tested Grades and Subjects. National Center for the Improvement of Educational Assessment. Retrieved from http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf, 4.

[69] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 9.

marginal impact on test scores across all students in a classroom, which doesn't account for differential learning or a teacher's ability to target instruction to individual students' needs.[70] There is also a need to disentangle treatment and pre-assignment variables to find the true teacher effects separated from past student outcomes.[71]

Other possible sources of error and bias come from the assumption in all value-added models of the random assignment of schools and teachers. If teachers and students were randomly assigned to communities, schools, and classrooms, achievement differences among classrooms would provide an unbiased ranking of teachers based on quality.[72] Random assignment is not common in most school districts or schools as principals typically influence classroom assignment, which affects the distribution of classroom average achievement levels within a school.[73] Another challenge in value-added modeling is in cases of few students. When a teacher has a small number of students, estimates of teacher effects can be heavily influenced by the performance of only a few students.[74] Researchers have found a much higher probability that quality estimates for school or teachers with small numbers of students will fall into the tails of the distribution, which is especially concerning as accountability systems that focus on those at the top or bottom are likely to disproportionately reward or punish low-

---

[70] Little, O., Goe, L., & Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/practicalGuide.pdf.

[71] Rothstein, J. (2008). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w14442.pdf.

[72] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 2.

[73] Ibid, 2.

[74] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 2.

enrollment schools or teachers.[75] Ultimately the desirability of any particular approach depends on how well it accounts for potential confounding factors to teacher quality.[76]

Another major empirical limitation in value-added modeling comes from the use of standardized tests, which may not fully capture all that students have learned or may be expected to know. Changes in the timing of tests, the weight given to alternative topics, or the methods used to create scores from students' response could also affect conclusions about the growth of achievement across classes of students.[77] Value-added measurement works best when students receive a single objective numeric test score on a continuous development scale, which is not necessarily tied to grade-specific content.[78] Most testing instruments sample items from a broader domain of skills, some of which may be more difficult to capture in a standardized test. This practice of sampling may leads to teacher narrowing of curriculum to standards most aligned to test.[79] Also, curricular differences among schools and districts may influence the time allocated to each subject and, therefore, knowledge of particular material.[80]

---

[75] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 4.
[76] Ibid, 3.
[77] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 3.
[78] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 14.
[79] Ibid, 16.
[80] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 4.

Other issues in using assessments include concerns with scaling, as many state assessments were created as a proficiency measures and therefore do not contain sufficient stretch for very low and very high achieving students. In a value-added model, measurement error in a prior test score used as a control variable biases the coefficient on the predicted current year test score. This random error in the test score leads to errors in ranking teachers and schools based on their true impact on knowledge measured by the tests.[81] The imprecision of value-added estimates does not imply that they have no productive uses, but rather may facilitate more informed uses of standardized test results and the development of stronger assessment.[82] Improving tests and adding items also makes prior achievement a better measure of accumulated knowledge, and researchers may also add other tests from previous years or other subjects as controls.[83]

Other important logistical considerations in the development and use of value-added measures include the accurate linkage of teachers and students and methods to account for missing student data. An accurate teacher-student link serves to identify who taught each student in each subject and for what percentage of instructional time. The teacher-student link is vital to ensuring correct value-added estimates and may not be straightforward in instances of team teaching or for students receiving supplemental ESL or special education services. The challenge of missing student data may create selection bias, as low achieving students are more likely to be absent or change schools during the school year. Although there are statistical methods to overcome these problems, states

---

[81] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 4.
[82] Ibid.
[83] Ibid, 5.

and districts must be careful to not systematically penalize or reward certain groups of teachers.

Policy Implications

In addition to many empirical concerns, the use of value-added data brings up many policy considerations such as deciding the variables to include in the model and how to explain complex results to relevant stakeholders. Value-added models must include separate teacher and school effects for each subject and each grade and describe how these affect all outcomes and persist over time. The model must also specify the correlation between teacher effects, school effects and residual error terms for different subjects within and across grades.[84] Isolating teacher and school effects can be difficult because of the need to account for uncontrolled factors that may be omitted or imperfectly measured.[85]

The most standard variable that is included in value-added models is a measure of lagged achievement. By controlling for a student's previous testing history, researchers can remove much of the variation in contemporaneous ability as well.[86] Variation in peer composition, class size, and other school characteristics remain are also likely to be systematically related to teacher quality. These additional factors in student achievement illustrate the value of using a multiple regression framework that uses information on family characteristics, class size and other school variables, and peer variables including

---

[84] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 73.
[85] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.
[86] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 3.

average lagged test score, racial composition, and turnover to control for remaining variation.[87] If variations in the composition of the school are not taken into account, these omitted variables may produce bias in applications of value-added measures.[88]

Another common policy concern with value-added models is the perceived statistical complexity and lack of transparency, or "black box" mechanisms, in many of these measures. Since these models rely on advanced statistical processes, teachers cannot calculate their own value-added estimates and may not understand how these results are found. Also due to some inherent statistical uncertainty, it is difficult to know the true effect size of an individual teacher in a single year. Given the complex relation between the many factors connected with student achievement, it is unlikely that a value-added regression will produce unbiased estimates of teacher fixed effects. The key issue is the magnitude of the imperfections.[89] Ultimately, acquiring a clearer understanding of the challenges faced in developing value-added measures allows for improvement in the methods used to estimate teacher value and can further inform how these estimates are used.[90]

COMPARING COMMONLY USED VALUE-ADDED MODELS

As a response to Race to the Top and the national focus on school and teacher quality, value-added models are being developed and used to inform teacher effectiveness

---

[87] Ibid.
[88] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 2.
[89] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 4.
[90] Ibid, 1.

across the country. The following section provides a framework for characterizing how different value-added model specifications differ in their assumptions and implications for conclusions on teacher effectiveness. The following sections detail the methodology, assumptions, limitations, and implementation of the Education Value-Added Assessment System, Student Growth Percentiles, Residual Models, and Hierarchical Linear Models.

Education Value-Added Assessment System (EVAAS)

The Education Value-Added Assessment System (EVAAS), developed by William Sanders in the 1990s, is one of the oldest value-added models in education.[91] The creators of EVAAS designed the model to "predict individual students' chances for success at future academic milestones."[92] The statistical process used by EVAAS allows for large-scale tracking of variation in student achievement test scores over time.[93] EVAAS is distinctive among value-added measures in that the model only uses prior test scores as predictors for current outcomes, without any controls for student, classroom, or school level characteristics. This model has limitations, including the need for multiple years of testing data. Additionally, it's been the target of criticism over both the perceived complexity of the model and the fact that the model design excludes student, classroom, and school variables in the model. EVAAS has been widely used in Tennessee (TVAAS), Ohio, and in the large urban school district profiled in Chapter 4.

---

[91] Sanders, W. (2000). Value-Added Assessment from Student Achievement Data: Opportunities and Hurdles. *Journal of Personnel Evaluation in Education, 14(*4).
[92] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf.
[93] Amrein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. Educational Researcher. Retrieved from http://edr.sagepub.com.ezproxy.lib.utexas.edu/content/37/2/65.full.pdf+html.

Methods

EVAAS uses different types of models according to the objectives of the analyses and the characteristics and availability of the test data.[94] The general type of model used in the analysis is a multivariate, longitudinal mixed model where the entire set of observed test scores for each student is fitted simultaneously.[95] This model is the best option when test scores are on a common scale. The univariate response model is an alternative EVAAS option in which student scores in a particular subject, grade, and year serve as the dependent variable and students' prior scores in multiple subjects, grades, and years serve as predictor variables.

Assumptions

The EVAAS model assumes that for each grade, the school effects, teacher effects, and the residual error terms are respectively independent and unbiased.[96] With multiple years of testing, the models typically assume that all cross-year correlation is explained by the inclusion of the prior years scores as a predictor variable and prior year teacher effects do not explicitly enter the model. The EVAAS model requires that standardized tests have the following psychometric qualities: "reliable, highly correlated with curricular objectives, and with sufficient stretch in the reporting scale to measure achievement of both very low and very high achieving students in a grade and subject."[97]

---

[94] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf.

[95] Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., et al. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 13.

[96] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 75.

[97] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf.

These qualities may not be the met with many state exams designed to measure student proficiency. EVAAS does not require complete data for each student and uses observed scores to predict missing student scores. This prediction of missing data reduces uncertainty in EVAAS estimates and minimizes selection bias through the inclusion of all students.

EVAAS also uses shrinkage estimation, which assumes that every teacher is average until the data shows otherwise.[98] This estimation technique may protect teachers from receiving an inaccurate estimate due to the accumulation of random errors, especially for classrooms with small numbers of students.[99] Shrinkage estimation may also increase the reliability of teacher effect estimates across the years, which will be explored in greater detail in chapter 3.

The most distinctive assumption in the EVAAS model is that a "student's testing history serves as his or her own control."[100] EVAAS includes a student's entire testing history over multiple years and subjects, but no socioeconomic or demographic data at the student, classroom, school, or community level. EVAAS has received criticism for not including a fuller range of variables, but the developers maintain that the prior achievement pattern of students contains all necessary information needed to make a student growth prediction. In *A Response to Criticisms of SAS EVAAS*, the EVAAS developers maintained that "the use of socio-economic status adjustments at the student level has largely been discouraged among statisticians and policy makers involved with

---

[98] Ibid.

[99] McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf, 75.

[100] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf.

value-added modeling, including the policies developed for Adequate Yearly Progress in growth model augmentations for No Child Left Behind."[101] Although other research has found a high correlation between average student achievement and percent minority and in poverty, EVAAS developers states that student achievement growth correlations with student characteristics vary from place to place but are "modest at worst and essentially zero at best."[102] The EVAAS developers also claim that adjustment for SES may over-adjust teacher estimates and "camouflage the fact that students in certain schools are not getting an equitable distribution of the teaching talent."[103] The inclusion of student, classroom, and campus demographics is a major distinguishing factor between value-added models, and this underlying assumption has a substantial influence in the findings in teacher effects.

Limitations

One of the limitations of EVAAS is that the model requires at least three prior student test scores to minimize selection bias and problems caused by errors of measurement in prior test scores. Since EVAAS doesn't include any student characteristics, the model uses all prior achievement test scores for each student in the predictor variable set. This inclusion means that prior reading, math, science, and social studies scores are all used to predict each of the current year's scores on a particular test. Without the inclusion of other school, teacher, and student characteristics, the EVAAS model assumes that these test scores explain all variation in student achievement patterns. One of the strongest critiques of the EVAAS model is its perceived "black-box" methods

---

[101] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf, 5.
[102] Ibid, 6.
[103] Ibid.

and lack of transparency. The EVAAS developers claim that less sophisticated approaches are more vulnerable to the problems of selection bias and increased uncertainty, which may over-identify very ineffective or very effective teachers and lack year-to-year reliability. This claim will be explored further in the following chapters in the discussion of model variability and persistence in value-added effects. The EVAAS developers claim to have created the model to prioritize reliability of analysis with a secondary focus on ease of interpretation and ease of usage.[104]

Another criticism of the EVAAS model is that it has not been peer reviewed. The developers of EVAAS claim that many types of linear mixed models are readily available and well understood by many other value-added modelers. A critique by Audrey Amrein-Beardsley echoes many of these concerns, and especially highlights that policymakers may be using the EVAAS model beyond how the model was originally intended. She elaborates that too few analyses have been conducted to examine and evaluate the validity of the inferences made in EVAAS value-added reports. She ultimately questions whether the EVAAS method will go beyond just reporting results to school to actually help to improve student learning.[105]

An additional critique of the EVAAS model is that the model's predictions of student performance aren't later verified with actual performance. EVAAS developers have responded to these concerns with the results from three states using EVAAS projection methodology, which are participating in the growth model pilot program of

---

[104] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf, 7.

[105] Amrein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. Educational Researcher. Retrieved from http://edr.sagepub.com.ezproxy.lib.utexas.edu/content/37/2/65.full.pdf+html.

NCLB. Through this project the EVAAS methodology was reviewed by four different peer review teams, and the analysis found that using prior test scores from multiple grades and subjects gave greater accuracy than predicting one year ahead using a single prior test score.[106]

As discussed in the previous section, the largest critique of EVAAS is the model's omission of student, classroom, and school characteristics as control variables. The developers of EVAAS have argued that the model implicitly controls for socioeconomic status and other background variables that are related to initial levels of achievement. Other education scholars question why the effects of important student characteristics variables should be completely accounted for in the prior year test score.[107] The exclusion of school effects also limits the models ability to disentangle school effects from teachers, which may lead to a biased estimate of teacher effects.[108]

Implementation

The Tennessee Value-Added Assessment System (TVAAS) is the first accountability system of its type to be adopted statewide.[109] The system was developed by William Sanders and colleagues and served as the basis for the development of EVAAS, which has been used in many other states and districts across the country. TVAAS has been used in Tennessee since 1993, and the primary purpose of the measure

---

[106] Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf, 8.
[107] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.
[108] Ibid.
[109] Millman, J. ed. (1997). *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Corwin Press, Inc., 133.

is to provide information about how effective a school, system, or teacher has been in leading students to achieve normal academic gain over a three-year period.[110] TVAAS uses student results on the Tennessee Comprehensive Assessment Program (TCAP) to measure student learning in grades 3 through 8 in science, math, social studies, language arts, and reading.[111]

The TVAAS reports on individual teacher effectiveness are made accessible only to administrators and teachers, although the general public has access to school and district-level EVAAS results. Tennessee's First to the Top Act required that EVAAS results be included as up to 50% of the evaluation system for teachers with this data. TVAAS results can be also used to create individualized professional development plans for teachers, which can be compared with later TVAAS results to judge the extent of improved teacher performance.[112]

Ohio has also developed a new accountability system involving multiple measures, including the EVAAS model. The Ohio accountability system is based on set of indicators that includes the percentage of students reaching proficiency on state tests, graduation and attendance rates, achievement of adequate yearly progress under NCLB, a performance index that combines state tests results, and a value-added indicator. The EVAAS model is being used as the value-added indicator, serving as a "customized prediction of each student's progress based on his or her academic record, as well as that of other students over multiple years, with statewide test performance as an anchor."[113]

---

[110] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 17.
[111] Millman, J. ed. (1997). *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Corwin Press, Inc., 137.
[112] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 18.
[113] Ibid, 19.

Ohio is only using EVAAS at the school level for elementary and middle schools and has not explicitly tied the results to teacher evaluations.

The de-identified large urban school district profiled in chapter 4 has used EVAAS since 2007 to calculate teacher and school-wide value-added for a district-wide performance-pay plan.[114] The district is now transitioning to use EVAAS for both performance-pay and high-stakes personnel decisions. Chapter 4 will provide a more in-depth analysis of the variability and implications of the use of EVAAS within this context.

Student Growth Percentiles

Student growth percentiles are another type of value-added model that estimate the distribution of students' current-year test scores given a history of prior-year test scores. In this method a group of students with a similar pattern of test scores are ranked into percentiles based on their performance on the current year test. This model uses a type of non-linear analysis in which each student's growth is compared to the growth of other students within the same quintile that allows growth to be assessed relative to a student's academic peers.[115] The student growth percentile model works by calculating the conditional percentile rank for each student's level of achievement on test Y compared to other students who had the same prior test score X. This model averages the gain for all students of a particular teacher, school, or district to obtain an indicator of effectiveness and then standardizes these results to a normal cumulative distribution

---

[114] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 12.

[115] Wright, P. S. (2010, March 10). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf, 1.

function.[116] In this model, a single prior year test score is the only necessary data, although the estimates are more precise with several years of prior test scores. Because of the comparison of students within similar peer groups, this model is generally more intuitive to understand than EVAAS.

Assumptions

Student growth percentiles generally have fewer assumptions than parametric models such as EVAAS. This model doesn't assume a normal distribution in the data or a linear relationship in student test scores.[117] Like all value-added models, student growth percentiles assume that teachers and students are randomly sorted, which is rarely the case in practice. SGP models are unique in that they don't rely on assumption of interval scaling on standardized exams and allow for transformation of the underlying test score scale.[118] Similar to EVAAS, student growth percentile models assume that prior student test scores are a complete proxy to predict student growth. These models don't include control variables that factor in student, classroom, or school characteristics. Relying on only prior test scores to make predictions assumes that students with the same pattern of test scores have those scores for systematic reasons and will continue to show similar patterns of growth, which may be violated in practice due to unobservable reasons.

---

[116] Ibid, 2.
[117] Wright, P. S. (2010, March 10). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf, 1.
[118] Briggs, D. & D. Betebenner. (2009). Is Growth in Student Achievement Scale Dependent? Retrieved from http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf, 3.

Limitations

Student growth percentiles have a major limitation in that these models can't provide standard errors with point-based estimates of teacher effectiveness. Teachers are scored based on the median percentile growth of students and not within a range or confidence interval. Without the ability to control for error in final value-added estimates for teachers, the accuracy of SGP calculations can't be measured. In some cases SGP models will be less precise than other value-added models since the model is dependent on calculating conditional percentiles for students with the exact same set of prior test scores on multiple tests.[119] Another limitation of SGP models is that they can't include control variables that factor in differences in student growth patterns that results from variation in student, class, and school characteristics.

Implementation

The most well-known student growth percentile model is the Colorado Growth Model. This model provides a common understanding of how individual students and groups of students progress from year to year toward state standards based on where each individual student begins.[120] The Colorado Growth Model allows the state to recognize schools and districts that produce the highest sustained rates of growth, regardless of their absolute test scores. Colorado developed the model to describe how much growth each student makes and how much growth is needed to reach state standards. It provides a complete history of all students' individual-level test scores from the Colorado Student Assessment Program (CSAP). The model also depicts academic growth in relation to

---

[119] Wright, P. S. (2010, March 10). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf, 2.

[120] Colorado Growth Model. (n.d.). Retrieved February 14, 2012, from http://www.cde.state.co.us/research/GrowthModel.html.

normative information about student progress toward the criteria of reaching different state proficiency levels.[121] The Colorado Growth Model has served as a framework for the development of many other student growth percentile models across the country.

The de-identified large urban school district profiled in chapter 4 is also piloting a version of a student growth percentile model, which the district is calling Comparative Growth. The district's Department of Research and Accountability uses the Stanford and APRENDA scores to calculate teacher's Comparative Growth rating. In this model, students are placed into a percentile group based on their previous year's test score and then ranked within their district-wide percentile group using their current year's scores. The district then calculates the median score for each teacher's students, which serves as the teacher's Comparative Growth score. A further analysis of this measure will be included in Chapter 4.

Residual Model

Another commonly used method for value-added analysis is through a residual model. The residual model for estimating value-added uses the statistical technique of regression analysis to predict average current-year test scores for students based on the students' prior-year test scores and other student, classroom, and school-level traits. The predicted score for each student is then compared to the student's actual score, and the residual difference between the two is considered the teacher's effect on student learning.[122] The value-added scores for all of the students of a teacher are averaged to

---

[121] Colorado Growth Model. (n.d.). Retrieved February 14, 2012, from http://www.cde.state.co.us/research/GrowthModel.html.

[122] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.

find the overall value-added score for each teacher. Residual models are generally more straightforward than EVAAS and allow for the inclusion of student, classroom, and school characteristics to better isolate teacher effects on student growth. The inclusion of multiple influencing factors on student achievement may allow for more precise estimates of teacher effectiveness. Residual models produce standard errors and confidence intervals that allow estimates of the precision of the results.[123]

Assumptions

Residual models use predicted current-year test scores as the dependent variable and prior-year test scores and a range of other variables as independent variables, including average classroom prior-year test scores and student, classroom, and school characteristics.[124] Residual models include the assumption that teachers and students are randomly sorted across schools and districts, which is generally not true. The lack of random sorting may bias the results, and small numbers of students may result in large standard errors in the value-added estimates. Residual models also assume that student assessment data are normally distributed, which is generally not the case with many state assessments.

Limitations

The major limitation of a residual model is that the model must be well specified to account for all variables that effect student learning besides the teacher. Possible unaccounted variables include students' prior knowledge and skills not captured in prior test scores, summer learning loss, and other immeasurable student background factors.[125]

---

[123] Ibid.
[124] Ibid.
[125] Ibid.

These non-included factors will result in error in value-added estimates since they are hidden in the teacher's results.

Implementation

The Gates Foundation Measures of Effective Teaching Project is using a residual model to calculate estimates of teacher effectiveness. The MET Project defines a teacher's value-added as "the mean difference, across all tested students in a classroom with a prior year achievement test score, between their actual and expected performance at the end of the year."[126] In this definition of value-added, if the average student in the classroom outperformed students elsewhere who had similar demographics and performance on last year's test and classmates with similar prior year test scores and other characteristics, the teacher is inferred to have contributed a positive achievement gain. This model is based on state tests and Stanford scores in Mathematics and English Language Arts in grades 4 through 8.[127] The project also uses two types of assessments that include cognitively demanding content, are well-aligned with the state curriculum, have high levels of reliability, and evidence of fairness to different groups of students.

The MET study is also unique in that it correlates the value-added achievement results with student perceptions, teacher observations, and past achievement results. The initial findings from the multi-year study have found that a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years. The study also found that student perceptions in one class are related to the achievement gains in other classes taught by the same teacher. The

---

[126] Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf, 10.
[127] Ibid, 11.

final goal of the project is to improve the quality of information about teacher effectiveness in order to help build fair and reliable systems for teacher observation and feedback.[128]

The District of Columbia Public Schools are also using a residual model that accounts for student characteristics that could be related to standardized test performance. The DCPS reports of teacher estimates also provide standard errors and confidence intervals associated with teacher value-added scores.[129]

Hierarchical Linear Models

Hierarchical linear models predict student achievement based on the nested relationships between students, classrooms, and schools. Many of these models include student prior achievement, student demographic characteristics, classroom-level characteristics, and school-level characteristics. The hierarchical linear model relies on a layered form of regression analysis to calculate value-added scores.[130] The model includes a level for student, classroom, and school variables and incorporates the effects from each level on the others.

Assumptions

Hierarchical linear models assume that there are underlying connections between the school, classroom, and student results that can't be fully accounted for in a simple

---

[128] Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

[129] Isenberg, E. & H. Hock. (2011). Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010-2011 School Year. Final Report. Retrieved from http://ddot.dc.gov/DCPS/Files/downloads/In-the-Classroom/Design%20of%20Value-Added%20Models%20for%20DCPS%202010-2011.pdf.

[130] Osborne, J. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=1.

covariate or residual model.[131] This model also allows for a full range of variables from the student, classroom, and school level in the model and is therefore extensive in its incorporation of relevant variables that may affect student performance. Hierarchical linear models also allow for correlation among predictor variables in the same level, which may help explain variance in the model.[132] Also, HLMs are more data intensive due to the large number of variables included in the model and are therefore harder to compute, especially with large numbers of students.

Limitations

One of the major limitations of hierarchical linear models is that they must be well specified and fully account for all factors that affect student learning besides the teacher. These models may also be difficult to use in districts with fewer schools and teachers within schools since the model relies on the nested relationship between these elements.[133] Also, HLM models rely heavily on the assumption of random assignment of students, which is relatively uncommon in schools and school districts.

Implementation

The state of Louisiana uses a hierarchical linear model to evaluate the quality of their educator preparation programs. Instead of assessing the effectiveness of individual teachers within the model, the state aggregates teacher effects to the preparation program level. This model allows the state to examine the efficacy of teacher preparation programs. In the first year of analysis, value-added scores were calculated for students in

---

[131] Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., et al. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 13.

[132] Osborne, J. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=1.

[133] Ibid.

Grades 4-9 in 66 of the 68 Louisiana public school districts. The results from this model allow for the separation of subject tests so that teacher effectiveness could be examined based on scores in the four tested subjects of English Language Arts, Mathematics, Science, and Social Studies.[134] The results from the evaluation found that the single largest predictor of student achievement was the student's prior test score in the content area, followed by prior achievement in other subject areas.[135]

[134] Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge, 26.
[135] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 41.

Chapter 3: Exploring the Variability and Stability of Value-Added Models

Although there are a wide variety of value-added models used to measure teacher effectiveness, the ultimate purpose of these models is to measure teacher effectiveness in a way that accurately captures true teacher effects each year and over time. The utility of value-added estimates of teachers' effects on student test scores depends on whether they can distinguish between high- and low-productivity teachers and predict future performance.[136] For any performance-based evaluation system to provide the correct incentives and enhance teacher quality, there must be a strong link between true performance and reward or retention.[137] Teacher effect estimates that exhibit low year-to-year correlations have limited utility because they fail to yield information that is sufficiently stable to support decisions about teachers.[138] Measures must provide accurate, unbiased measures of teacher productivity to assure the measures' efficacy in high-stakes personnel decisions. If value-added measures vary substantially over time, a tenure policy based on a short time frame could lead to the dismissal of many truly effective teachers and the retention of others who prove to be relatively ineffective in boosting achievement.[139] Exploring the reliability and stability of value-added models over time is vital to reaching true estimates of teacher effects.

---

[136] McCaffrey, D., Sass, T., Lockwood, J.R., & K. Mihaly. (2009). The Intertemporal Variability of Teacher Effect Estimates. *American Education Finance Association*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2009.4.4.572.
[137] Ibid, 573.
[138] Ibid, 579.
[139] Ibid, 573.

EXAMINING CORRELATIONS IN TEACHER RESULTS

One of the challenges in setting expectations for consistency in value-added measures is that few studies have measured the variability of teacher effects over time and between tests. To reach a closer understanding of true teacher effects, researchers need to examine the variability of teacher effect estimates obtained using alternative models or using data from the same teachers over time or across different course offerings.[140] Few research studies have looked at the long-run persistence of teacher effects on achievement.[141] The few research studies since the 1950s has found a substantial amount of variability in teacher effects over time. Rosenshine's work found year-to-year correlations over teacher effect only as high as 0.5, and average correlations were about 0.35 or lower.[142]

There are also major variations in findings with different tests and statistical models. Much of the empirical work addressing the consistency of teacher effectiveness over time is inconclusive.[143] Another major challenge is that most states and districts contract with a single vendor for value-added measures and therefore have no reference point to compare model results. Reaching an understanding of the source of year-to-year

---

[140] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.

[141] Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA.

[142] Rosenshine, B. (1970). The Stability of Teacher Effects Upon Student Achievement. *Review of Educational Research*. Retrieved from http://rer.sagepub.com/content/40/5/647.

[143] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810, 7.

variability will have implications on how to best use the effectiveness measures for evaluating teachers.[144]

Although correlation and consistency measures are weakened by the challenges of non-random assignment of students and test reliability, there is also a need to examine consistency in teacher behaviors over time to isolate a true teacher effect. With clearer information on the variables that influence teacher consistency, it will be easier estimating the stability of coefficients that might be expected in different situations.

The expected amount of variance from year to year will have a major impact on both the utility and design of value-added measures, especially in relation to evaluation decisions for teachers. Relatively low intertemporal correlations for teachers may not be out of line with findings from other occupations that measure productivity more directly.[145] For example, researchers have found that volatility in teacher's value-added between years is no higher than for performance measures used in Major League Baseball. Smith and Schall found that the between-season correlation in batting averages was 0.36, and the between-season correlation for major league pitchers was 0.31.[146] The time frame for measuring performance will significantly influence the findings. Over time, correlations will decline as seen in the performance of salespersons, university faculty, and baseball players.[147]

---

[144] McCaffrey, D., Sass, T., Lockwood, J.R., & K. Mihaly. (2009). The Intertemporal Variability of Teacher Effect Estimates. *American Education Finance Association*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2009.4.4.572, 573.
[145] Ibid, 593.
[146] Smith, G. & Schall, T. (2000). Do baseball players regress toward the mean? *The American Statistician, 54,* 231-245.
[147] Ibid.

Although correlations in teacher results may decrease over time, value-added results are one of the strongest predictors of future student achievement for a teacher. The research finds that after only one or two years of student outcome data, a district has important additional data about which teachers are likely to generate large student learning gains and which are not. Value-added measures provide stronger information on the tails of the distribution (for the most effective and least ineffective) than for the majority of teachers who are in the middle.[148]

RESEARCH FINDINGS ON VALUE-ADDED VARIABILITY

The following sections highlight research findings from four case studies that have examined variability in value-added models that assess teacher effectiveness. This section includes research examining the choice of outcome measures, using different value-added models, and the stability of teacher effects over time. These case studies highlight the core research questions about the variability in value-added results, findings in the persistence of teacher value-added measures, and the implications of the results. The research studies set the stage for an analysis of the persistency and variability of a large urban school district's value-added results in Chapter 4.

Variability in Teacher Effectiveness based on Choice of Outcome Measure

Using data from the Houston Independent School District, Corcoran, Jennings, and Beveridge examined how the choice of outcome measure affects inferences about teacher quality.[149] This question is largely unexplored in the research and has important implications for what sorts of measures are used in value-added models. This study seeks

---

[148] Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution, 9.

[149] Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA.

to fill the research gap by contributing a theory to explain the wide variation in teacher effectiveness that cannot be well explained by traditional measures of quality, such as years of teaching experience. The research also seeks to examine the "tacit assumption of value-added systems that these measures are meaningful, reliable, and relatively stable indicators of teaching effectiveness."[150]

The team used data from the Houston Independent School District to estimate teacher effects on high and low-stakes tests of the same content areas. The researchers compiled a longitudinal dataset of all students tested in Houston between 1998 and 2006, approximately 165,000 students per year. The study estimates teacher effects using an identical sample of students and up to eight years of classroom data for each teacher. In order to compare the teacher effects across tests, the study includes student results from both the Texas state assessments (TAAS or TAKS) and the Stanford Achievement Test (SAT) Battery. The study limited their sample to 4th and 5th grade math and reading scores to provide a lagged achievement score and ensure correct linkage for students to teachers.

The research study found that teachers' effects are 15-31% larger on the high-stakes test and that teacher effects on the high-stakes test are only a modest predictor of effectiveness on the low-stakes test. The study also found that returns to experience differ across tests in ways consistent with teachers' incentives to invest early in teaching skills and content specific to the high-stakes test. In their analysis of persistence, the researchers found that teacher effects on the high-stakes test decay at a faster rate than those on the low-stakes test. In overall teacher effect, the team found large effects of 4th and 5th grade teachers on achievement in both reading and math, with a single standard

[150] Ibid, 4.

deviation increase in teacher effectiveness associated with a 0.205 standard deviation increase in reading achievement and 0.256 standard deviation increase in math. The team found that overall magnitude of teacher effects varies with the test and that there was greater variation on the high-stakes test than on a low-stakes test of the same subject. The teacher effects on the high-stakes reading test were 18-31% larger than the low-stakes test and 15-26% larger on the high-stakes math test.

In the analysis of the correlation between results from the high stakes and low stakes tests, the researchers found the correlation in teacher effects between the TAAS/TAKS and SAT as 0.499 in reading and 0.587 in math. The study also found that the correlation in teacher effects is much stronger between subject areas on the same test (0.675 for the TAAS/TAKS and 0.625 on the SAT) than across tests of the same content area. These correlations yield inconsistent rankings of teachers, especially when teachers only have a single year of results. In the analysis of the quintile rankings, only 43% of those in top quintile on TAAS/TAKS reading were also into the top quintile on the SAT and 17% were in the bottom two quintiles. A threshold for exceptionally low or high performers (the top or bottom 5-10%) would have few teachers, especially when measured across all 4 tests. The study found that 1.6% of teachers ranked in the bottom decile of all four tests, and only 0.4% ranked in the bottom 5% of all four tests. They also found that only 28% of those in the bottom 5% also ranked in the bottom 5% of the other test in the same subject.

These inconsistencies in teacher effects across tests resemble the pattern of year-to-year variation in teacher effects also found in other research studies. The study also found notable differences in the returns to teaching experience across the two tests, where more experienced teachers had a greater effect on the SAT measure than less experienced teachers. This analysis of the magnitude of the effects and the relative teacher rankings as

implied by each test paints a picture of relative inconsistency between the two measures, which calls into question the validity of these measures, especially in high-stakes decisions.

The researchers explore a wide variety of plausible reasons that teacher effects might vary across tests. Student and classroom-level noise both contribute to inaccuracies in estimates of true achievement and control for all outside factors. Differences in tested populations where students may be excluded from the high-stakes test may also contribute to variations in teacher effects across tests. Another major variable is test content and difficulty since state tests emphasize state curriculum, and national tests draw from a broader domain. Student effort may be another important variable in the analysis of the results since students' investment in a test may vary depending on the incentive to perform well. This hypothesis was also supported in the findings that the correlation between student scores on the two tests was more highly correlated on the SAT, which is a low-stakes test. Another major factor influencing achievement is the teacher incentives tied to the test, which influence teacher behavior. Teachers and schools are specifically rewarded for increasing TAKS scores, not the broader set of skills that are captured on the SAT. Also the TAKS results are tied into teacher evaluations, especially at the beginning of a teacher's career.

The implications of this research show that the choice of outcome measure has a major impact of the conclusions draw about teacher effectiveness. If the estimates of teacher effects could be taken as causal effects on student achievement, the high- and low-stakes test would offer different conclusions about the relative contribution of teachers to test scores. The SAT implies a 20% smaller impact of teacher quality on

achievement.[151] The research study also concludes that test-based accountability may incentivize teachers to focus efforts on short-term, test-specific skills that may not generalize to other tests. The researchers also emphasize that the results do not suggest that one test is superior to another for constructing value-added measures or that an estimate that combines results from the two tests would be an unambiguous improvement over a single test battery. Ultimately this research shows the variability in value-added measures based on the accountability measures tied to the assessment and results.

Initial Findings from the Measures of Effective Teaching Project

The Measures of Effective Teaching Project was created based on three simple premises: whenever feasible, a teacher's evaluation should include his or her students' achievement gains, any additional components of the evaluation should be demonstrably related to student achievement gains, and the measure should include feedback on specific aspects of a teacher's practice to support teacher growth and development. The project is measuring student achievement based on existing state assessments and with three supplemental assessments designed to assess higher-order conceptual understanding. Similar to the research by Corcoran, Jennings, and Beveridge, the study seeks to identify teacher effectiveness based on results on both high and low stakes assessments.

Although the MET project is only in the beginning stages, the research team has begun a preliminary analysis of the first year results. In the team's initial findings, they found that the correlation between a teacher's value-added on the state test and their value-added on the Balanced Assessment in Math was .377 in the same section and .161

[151] Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA, 25.

between sections of the test. To calculate the true correlation in teacher effects between these assessments, the study compared the results on the state math tests and the Balanced Assessment in Math with two different groups of students. This comparison estimated the correlation between the persistent component of teacher impacts on the state test and on BAM to be .54. These results imply that teachers with strong value-added are not simply "teaching to the test" to inflate student achievement, but are enhancing long-term conceptual knowledge. The initial findings in the correlation between the persistent component of teacher impacts the ELA state tests and Stanford 9 OE was .37, although recent changes in the NYC tests may have overly influenced this result.[152]

In the study's analysis of the correlations in teacher results, the team found similarly low correlations as Corcoran, Jennings, and Beveridge's research. The between-year correlations in teacher value-added were below 0.5, which implies that more than half of the observed variation is due to transitory effects rather than stable differences between teachers. The project observed the highest correlations in teacher value-added on the state math tests, with a between-section correlation of .38 and a between-year correlation of .40. The correlation in value-added on the open-ended version of Stanford 9 was .35. The correlation in teacher value-added on the state ELA test was .18 between sections and .20 between years. These correlations report the proportion of the variance that is due to persistent differences between teachers, which is still quite large given the range of total unadjusted variance in teacher value-added.[153] Similar to the results from

---

[152] Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf, 21.
[153] Ibid, 19.

Corcoran, Jennings, and Beveridge, this study also shows that the correlations in teacher value-added scores are relatively low over time and may be too unreliable for use in high stakes decision making in teacher evaluation.

An Exploration of Value-Added Stability across Models and Contexts

In the quest to further explore value-added models for measuring teacher effectiveness, Newton, Darling-Hammond, Haertel, and Thomas examine stability of high school teacher effectiveness rankings across differing conditions.[154] This study specifically aims to fill the gap in research examining the variability of teacher effect estimates obtained using alternative models of using data from the same teachers over time or across different course offerings. This research also seeks to examine the assumption of large, stable teacher effects, which most value-added models rely on for validity. Through an empirical investigation of the stability of teacher effectiveness ratings based on value-added modeling, this study examines the key assumptions of value-added models and the implications of using these measures, especially in high-stakes teacher evaluation decisions.

This study used a sample of 250 secondary teachers and roughly 3500 students taught by these teachers and specifically looks at the results of teacher effectiveness across statistical models, classes taught, and year. The researchers used English Language Arts and Mathematics courses for their analysis due to the overlapping constructs and skills from year to year. The study based measurement of value-added on the variation in pupils' test scores on the California Standards Tests (CSTs) controlling for prior-year

---

[154] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.

scores. The researchers used multiple models to control for key demographic variables, school fixed effects, and to account for students nested within classrooms and teachers nested within schools. Teacher effectiveness was measured by the average difference between actual and predicted scores for all students assigned to that teacher.

The research analysis was designed to investigate whether teacher rankings were consistent across different models, across different courses for teacher who taught multiple types of ELA or math courses, and across two year for teachers with three years of student test scores. The study found that teacher ratings were highly correlated with one another in both Mathematics and English Language. The teacher rankings inter-year correlations were modest (0.4 for ELA teachers and around 0.6 for math teachers) and fluctuated across models, courses, and years. The study also found that 74-93% of teachers' rankings changed by 1 or more deciles across years.

This research study also analyzed the impact of student characteristics on the variability in teacher ratings. The study found that student characteristics dramatically impact teacher rankings, even when characteristics are controlled for in the model. In this study, teacher with less advantaged students typically received lower effectiveness ratings than the same teacher teaching more advantaged students in a different year. The research also found that even models that accounted for student demographics showed negative correlations with the proportions of students who were English language learners, free lunch recipients, or Hispanic. The study also found that prior student achievement and the assignment to a high track vs. low track course were greater predictors of test scores than the teacher.

This research highlights the inherent difficulty in developing a value-added model to capture teacher effectiveness when teacher effectiveness itself is a variable with high levels of instability across contexts. These findings challenge the value-added measures

assumptions that teacher effects are a fixed construct independent of the context of teaching and stable over time. Since judgments of teacher effectiveness can vary substantially across statistical models, classes taught, and years, these measures must be carefully constructed and evaluated for use in high-stakes teacher accountability.

The Intertemporal Variability of Teacher Effect Estimates

A research study by McCaffrey, Sass, Lockwood, and Mihaly examines the year-to-year variability in estimates of teacher effects from value-added measures. This research study is based on the underlying premise that the utility of value-added estimates of teachers' effects depends on whether they can distinguish between high- and low-productivity teachers and predict future teacher performance.[155] The study also seeks to examine the implications of incentive policies based on a short time frame when value-added effects may vary greatly over time. This research specifically examines with within-teacher variance in estimated teacher effectiveness over time and the associated implications for a viable outcome-based system of teacher personnel decisions.

To test the variance in teacher effects over time, the study uses data from elementary and middle school mathematics teachers from five large Florida school districts. This study specifically sought to decompose the variance of teacher effects to provide insights into the relative utility of alternative achievement model specifications for estimating teacher effects. The study identified two key sources of variation over time in the annual teacher effect estimates: sampling error and nonpersistent changes in performance. This research also estimated variance components to characterize the various estimators of teacher effects as measures of teacher performance. Through the

[155] McCaffrey, D., Sass, T., Lockwood, J.R., & K. Mihaly. (2009). The Intertemporal Variability of Teacher Effect Estimates. *American Education Finance Association*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2009.4.4.572.

use of reliability and stability coefficients, the study is able to differentiate teacher performance in a given year and examine the proportion of variability in estimates that is due to persistent effects.

This study also describes the degree to which individual teacher estimates vary over time in comparison to the measured performance of workers in other occupations. They also determine the degree to which the within-teacher variance can be explained by observable time-varying teacher characteristics such as experience, formal education attainment, and in-service training. The study also explores the effect of averaging teacher effect estimates over multiple years and the implications of using single-year or multiyear estimates of teacher effectiveness in practical systems of teacher evaluation. The study uses results from both exams given by the state of Florida: the Sunshine State Standards Florida Comprehensive Achievement Test, which is criterion-based, high-stakes test designed to assess the skills that students are expected to master at each grade level, and the FCAT Norm-Referenced Test, which is a version of the Stanford Achievement Test. Similar to the results from Corcoran's research and the initial MET findings, the team found that using gains from one test or the other did not lead to consistent differences in year-to-year correlations of teacher effectiveness, but that using different tests can affect the stability of estimated teacher effects.

The study found that year-to-year correlations in value-added measures in the range of 0.2-0.5 for elementary school teachers and 0.3-0.7 for middle school teachers. The researchers found that teacher rankings have only moderate stability, where roughly one-third of top-quintile teachers remain in the top quintile the next year, while approximately one in ten falls to the bottom quintile of the teacher effectiveness distribution. The research concludes that roughly 30-60 percent of variation in measured teacher performance is due to sample error from "noise" in student test scores. They also

found that little of the variation in a teacher's performance over time can be explained by observable teacher characteristics like experience, attainment of advanced degrees, or in-service training. The study also found that averaging estimates from two years reduces sampling error and increases the ability to predict future teacher performance by roughly 50 percent. This research also found that using student fixed effects in models of achievement gains rather than unchanging student characteristics like race and gender increases sampling error in estimated teacher effects.

The core implication from this research is that it is difficult to control for bias and stability of measures over time. Attempts to reduce bias can come at the cost of lower stability estimates, while too little effort to remove bias can yield estimates that are unduly stable across years. The research concludes that if a district were to retain only teachers in the top three quintiles of distribution of true effectiveness, the average effectiveness of teachers would improve by about 0.04 of a standard deviation unit of student test scores. The inherent instability in value-added measures over time leads to caution about the use of these measures in high-stakes decisions, especially with a single year of results. The study also suggests that more qualitative measures may serve as a complement to VAM in evaluating teachers to increase reliability in assessments of teacher effectiveness.

IMPLICATIONS OF VARIABILITY IN TEACHER RESULTS

Each of the four reviewed case studies found instability in teacher value-added results over time, which has important implications for the validity of these measures in assessing teacher effectiveness. Another important study exploring the validity of teacher value-added results is Jesse Rothstein's falsification test.[156] This study explores the value-

[156] Rothstein, J. (2008). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *National Bureau of Economic Research*. Retrieved from

added assumptions about the nature of the educational production function and the assignment of students to classrooms. If these core assumptions are violated, the estimates of teachers' causal effects will be systematically biased. As opposed to random error, bias systematically penalizes the same group of teachers. To explore this assumption, Rothstein developed a falsification test for three widely used VAM specifications, based on the idea that future teachers cannot influence student's past achievement. His research finds that students who did poorly in 4th grade will predictably post unusually high 5th grade gains as they revert toward their long-run means. This regression to the mean led to statistically impossible large effects of 5th grade teachers on 4th grade test score gains.

This research shows that conventional measures of individual teachers' value added may fade out very quickly and are only weakly related to long-run effects. His research also found that a teacher's effect in a single year of exposure is correlated only 0.3 to 0.5 with her cumulative effect over two years and correlations with three-year cumulative effects are around 0.4. He also found a lot of movement between quintiles, as the fraction of teachers in the top and bottom quintile who were assigned the same quintile on another model were around 0.43 for math and 0.35 for reading. The findings from Rothstein's work and other researchers show that value-added measures need to be evaluated for model assumptions, stability and persistence over time, and the accuracy of results, especially in high stakes situations.

EXPECTATIONS FOR CORRELATIONS OVER TIME

---

http://www.nber.org/papers/w14442.pdf.

As seen in the reviewed research, expectations for consistency of teacher value-added measures over time may vary considerably based on the type of test, noise in the measurements, and natural variation in performance from year to year. Multiple studies have also found that, of teachers who ranked in the top 20 percent of effectiveness one year, less than a third of those had scores in the top 20 percent the next year, though the vast majority stayed in the top half.[157] Although rankings based on value-added estimates change from year to year, some of that change doesn't necessarily reflect an actual change in teacher effectiveness.[158] With the high degree of variability between teacher value-added results over time, it is essential to continually examine the connections between school, teacher, and student outcomes each year and over time.[159]

---

[157] Sass, T. (2008). The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy. National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/uploadedpdf/1001266_stabilityofvalue.pdf.

[158] Hull, J. (2011). Building a Better Evaluation System: Full Report. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Building-A-Better-Evaluation-System/Building-A-Better-Evaluation-System.html

[159] Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf, 5.

Table 1: Research Findings on Correlations in Teacher Effects

Researchers (Year)	Type of Correlation	Correlation estimates
Rosenshine (1970)	Teacher effects from year-to-year	Highest 0.5, Average 0.35 or lower
Corcoran, Jennings, & Beveridge (2010)	Teacher effects between tests (TAKS/TAAS & SAT)	Reading 0.499 Math 0.587
	Teacher effects between subject areas on the same test	TAAS/TAKS 0.675 SAT 0.625
Measures of Effective Teaching Project (2010)	Teachers value-added between tests (Balanced Assessment in Math & state assessment)	In the same section 0.377 Between sections 0.161
	Persistent component of teacher impacts on the math state test and Balanced Assessment in Math	0.54
	Persistent component of teacher impacts on the ELA state test and Stanford 9 OE	0.37
	Teacher value-added between years on state tests	Math 0.40 Reading 0.20
Newton, Darling-Hammond, Haertel, & Thomas (2010)	Teaching rankings across years	ELA 0.4 Math 0.6
McCaffrey, Sass, Lockwood, & Mihaly (2009)	Year-to-year value-added measures for teachers	Elementary school 0.3-0.7 Middle school 0.2-0.5
Rothstein (2008)	Teacher effect in a single year to cumulative effect over 2 years	0.3-0.5
	Teacher effect in a single year to cumulative effect over 2 years	0.4
Smith & Schall (2000)	Between season Major League Baseball batting averages	0.36
	Between season Major League Baseball pitching averages	0.31

Chapter 4: A Case Study Using Value-Added & Student Growth Models

In light of the research on the variability and stability of value-added models, this chapter uses data from a de-identified large urban school district to explore these trends in a case study and explore the implications of these findings. This district has used value-added modeling since 2006 to help draw conclusions about teacher effectiveness and is expanding the use of value-added models in making high stakes decisions. Beginning in 2012-13 the district will be one of the first to use both a value-added and alternative student growth method as data points to measure a teacher's impact on student achievement. This chapter explores the background on the use of value-added measures in the district, the research methodology used for estimates of variability of and between value-added and student growth measures, correlation results from the district's data, and the implications of these results. Ultimately this chapter continues the discussion from the previous section in calling for careful examination of the results from these measures over time to ensure accurate conclusions about teacher effectiveness.

Background on Value-Added Measures

In the 2005-2006 school year, this large urban school district first began a district-wide performance pay system based on an in-house calculation of teacher effects. After the district experienced some challenges with this original system, the district contracted with SAS to calculate value-added scores for core content teachers and school-wide value-added. These SAS EVAAS calculated value-added scores have been used for the district's performance-pay plan since 2007 and will be used for high-stakes decisions in teacher evaluation beginning in the 2012-13 school year.

The original performance-pay strands were based on a combination of school-level awards, individual teacher awards for those whose students' progress ranked in the top two quartiles for their grade and subject, and a mix of additional bonus opportunities, including attendance.[160] The maximum bonus can range from $6,600 to $10,300 for classroom teachers. Almost 90% of eligible school employees received a bonus for 2008-2009, and classroom teachers earned an average of $3,606.

This district's value-added results are generated by SAS EVAAS from the combined results on the Texas Assessment of Knowledge and Skills (TAKS), the Stanford 10 Achievement Test (or the Aprenda, the Spanish language equivalent), and multiple years of test results to calculate teachers' cumulative value-added. The expected scores in each year are estimated for students in each subject and compared with their actual scores. The value-added model only includes prior test scores as a complete control for student background characteristics. The EVAAS results also use the Texas 2006 state results as a benchmark for student progress.

Although this district has worked with SAS EVAAS since 2006 for results for their performance-pay system, the value-added results will soon be a major component of teacher evaluation in the district. In May 2011, the board of education voted to approve the use of value-added measures in the district's new teacher evaluation system, although the implementation of student achievement component of the new evaluation system was delayed for a year due to the new STAAR test. In a letter to the school board supporting the use of value-added measures, the superintendent of this district expressed a willingness to create "a screening process for principals who propose that teachers gain

[160] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 12.

term contract by requiring them to discuss the performance/effectiveness of all probationary teachers. This discussion will include the review of value-added."[161] The district is also contracting with Battelle for Kids to provide training for teachers on the instructional relevance of value-added information.

Proposed Alternative Student-Growth Model

In addition to EVAAS, the district is developing an alternative student-growth model as a complement the teacher value-added. This measure is similar to the Colorado Growth Model and is designed to examine the extent to which students grow as determined by benchmark scores for similarly performing students. This district's Department of Research and Accountability uses the Stanford and APRENDA scores to calculate teacher's Comparative Growth rating. In this model, students are placed into a percentile group based on their previous year's test score and then ranked within their district-wide percentile group using their current year's scores. The district then calculates the median score for each teacher's students, which serves as the teacher's Comparative Growth score. The district will begin training on the Comparative Growth measure in the summer of 2012, and the Comparative Growth component will be implemented with the other student achievement measures in the 2012-13 school year.

RESEARCH METHODOLOGY

Motivated by the limited body of research on the stability of value-added and student growth models, the primary research goal of this study is to examine the consistency of value-added and student growth models between subjects, over time, and

[161] Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf, 13.

across models and examine the implications of using these models in evaluations of teacher effectiveness. The core hypothesis is that using value-added and student growth models for student achievement will not necessarily give stronger conclusions about teacher effectiveness since these measures are not highly consistent between subjects, over time, and across models. As explored in the previous chapter, the research shows low to moderate correlations of teacher effects over time, and these measures may not show consistent teacher effects between subjects and across models.

To examine the stability of value-added results over time, this research study examines the correlations of the EVAAS results for teachers from this large urban school district between tested subjects across five years of results. After examining the stability of EVAAS results over time, the case study also analyses the correlations in the results from the EVAAS and Comparative Growth measure across subject and grades for a single year of data. The research study also examined the correlations in quartile rankings of teacher effects, although there were no significant differences in this analysis than with the initial correlations. Through this analysis, the district will have a stronger understanding of the consistency and reliability of EVAAS results between subjects, across years, and in comparison to the Comparative Growth results. The study concludes with the possible implications of these findings, especially in high-stakes teacher evaluation decisions.

The data used to conduct this analysis includes the EVAAS results for approximately 4,000 teachers per year from the 2006-2007 to 2010-2011 school years and 3,134 teachers in a matched set of data with both an EVAAS and Comparative Growth score for 2010-11. Since 2010-11 was the first year that a Comparative Growth score was calculated for teachers within this district, these calculations are considered preliminary and were only used to shape the development of the finalized model for

2012-13. Also, the 2006-2007 EVAAS data had some challenges in the first year of confirming the student to teacher link and are therefore not as precise at indicating teacher effectiveness as later years of EVAAS data.

The data are presented as an EVAAS cumulative gain index for each teacher for the following subjects: Language Arts, Math, Reading, Science, and Social Studies. The Comparative Growth results represent a teacher percentile for each grade and subject ranging from 0 to 99. To merge the results across the years, the data was compiled across years resulting in approximately 6,500 unique teacher IDs. Tables 2 and 3 below provide the descriptive statistics for the data.

Table 2: Descriptive Statistics for EVAAS Value-Added Results

Subject and Year	# of Obs	Mean	Std. Dev.	Min	Max
Language 2006	2028	0.06	1.59	-13.16	7.12
Math 2006	1886	-0.07	2.17	-.9.52	9.34
Reading 2006	2040	0.01	1.34	-8.47	7.6
Science 2006	1216	-0.02	1.65	-7.16	9.4
Social Studies 2006	1171	0.03	1.63	-7.75	11.01
Language 2007	2078	0.04	1.52	-9.56	9.57
Math 2007	1888	0.03	2.18	-9.34	12.1
Reading 2007	1975	0.01	1.41	-9.84	9.62
Science 2007	1237	0.03	2.24	-8.06	16.19
Social Studies 2007	1315	0.13	2.35	-9.37	7.85
Language 2008	1836	0.05	1.77	-6.41	9.75
Math 2008	1961	0.03	2.34	-12.63	9.11
Reading 2008	2046	0.01	1.52	-7.13	11.15
Science 2008	1286	0.02	-2.16	-10.1	16.64
Social Studies 2008	1335	0.03	2.26	-11.9	13.77
Language 2009	2101	0.07	1.82	-7.57	8.27
Math 2009	1949	0.11	2.44	-11.53	11.62
Reading 2009	1947	-0.01	1.63	-6.65	6.51
Science 2009	1300	0.02	2.34	-11.57	18.45
Social Studies 2009	1311	0.02	2.33	-10.36	14.03
Language 2010	2051	0.11	1.84	-7.1	7.69
Math 2010	1891	0.11	2.39	-12.48	11.53
Reading 2010	1910	0.03	1.51	-6.29	7.34
Science 2010	1237	0.11	2.19	-8.91	11.21
Social Studies 2010	1294	0.13	2.44	-7.64	14.94

Table 3: Descriptive Statistics for EVAAS and Comparative Growth Results in Matched Set

Subject, Grade, Model Type	# of Obs	Mean	Std. Dev.	Min	Max
Math 3rd CG	599	51.30	20.03	7	99
Math 3rd EVAAS	599	0.09	2.13	-6.5	8.59
Read 3rd CG	598	50.10	17.97	4	96
Read 3rd EVAAS	598	-0.01	1.56	-6.29	5.72
Math 4th CG	513	52.11	19.36	0	99
Math 4th EVAAS	513	0.06	2.26	-12.48	7.6
Read 4th CG	513	50.55	17.68	2	97
Read 4th EVAAS	513	-.01	1.43	-5.1	7.34
Math 5th CG	358	50.59	18.14	8	97
Math 5th EVAAS	358	0.13	2.35	-9.37	7.85
Read 5th CG	374	50.12	16.38	8	95
Read 5th EVAAS	374	0.04	1.66	-4.48	7.29
Science 5th CG	377	50.74	15.76	9	95
Science 5th EVAAS	377	-0.01	2.08	-6.47	7.11
Math 6th CG	154	52.84	15.22	16	89
Math 6th EVAAS	154	0.18	3.36	-9.68	9.6
Read 6th CG	175	49.10	12.99	12	83
Read 6th EVAAS	175	0.19	1.59	-4.35	4.21
Math 7th CG	148	52.44	13.67	22	87
Math 7th EVAAS	148	0.21	2.48	-5.55	11.53
Read 7th CG	147	50.95	10.60	19	75
Read 7th EVAAS	147	0.07	1.21	-2.86	3.06
Math 8th CG	150	51.74	13.59	12	92
Math 8th EVAAS	150	0.37	2.58	-7.28	6.89
Read 8th CG	139	50.30	10.81	14	75
Read 8th EVAAS	139	0.15	1.11	-2.23	4.15
Science 8th CG	130	48.52	12.23	19	82
Science 8th EVAAS	130	.28	2.84	-5.51	8.65
Soc. Stud. 8th CG	103	49.27	10.65	27	78
Soc. Stud. 8th EVAAS	103	0.63	3.34	-7.64	8.33

To find the correlations for the analysis, pairwise correlations were found between subjects and across years with the EVAAS longitudinal data and between models by grade and subject with the Comparative Growth and EVAAS matched set. The number of teachers included in the analysis is noted beneath the correlation results in each table. Teachers were only included in the pairwise correlation if they had results for both items being compared. The analysis of the correlations in quartile rankings found comparable results to the pairwise correlations due to the standardized nature of both the EVAAS and Comparative Growth results.

RESULTS OF ANALYSIS

The overall findings from the analysis found that the correlations between EVAAS results over years, across subjects, and with the Comparative Growth model ranged from 0.087 to 0.607. The statistically significant correlations in EVAAS results by subject across the five years in the data set ranged from 0.12 in science across 3 years to 0.454 in social studies across 2 years. The average correlation by subject over the five years ranged between 0.221-0.370. The correlations in EVAAS result across subjects in a single year ranged from 0.087 between Reading and Science in 2007 to 0.559 between Science and Social Studies in 2009. The average correlation by year over the five subjects rose across the five years in the sample starting at 0.199 in 2006 and increasing to 0.373 in 2010. The statistically significant correlations between the Comparative Growth and EVAAS results ranged from the highest value of 0.607 in 6th grade Math and the lowest value in 7th grade Reading of 0.243. All of the correlations in EVAAS results over years, across subjects, and with the Comparative Growth model were significant at $p<0.01$ level unless otherwise noted.

The results are listed in more detail in the following tables and sections, which compare the correlations in the EVAAS results in a single subject over the five years of data (Tables 4-8), the EVAAS results in a single year across subjects (Tables 9-13), the overall correlations by grade and subject between the EVAAS and Comparative Growth results (Table 14), and the EVAAS and Comparative Growth results by the grade level (Tables 15-17). Compared with the research results seen in Table 1, these results are somewhat similar and in some cases higher than the findings from other studies of correlations in teacher value-added scores over time. The implications of these results are discussed in more detail at the end of this chapter.

Correlations between EVAAS results within a subject across years

As seen in tables 4-8, within a single subject, the correlations in the EVAAS results across the five years of analysis ranged between 0.12-0.454. These correlations were lower than hypothesized, as one would expect the relative teacher effect in a single subject to be fairly persistent over time. The correlations within a single subject were highest in social studies and science, but still average around 0.3 and diminish over time. All of the correlations in this analysis were statistically significant.

Table 4: Correlation between EVAAS Language Arts Results Across Years

	2006	**2007**	**2008**	**2009**	**2010**
2006	**1.000**				
# of Teachers	2028				
2007	**0.244**	**1.000**			
# of Teachers	1466	2078			
2008	**0.209**	**0.347**	**1.000**		
# of Teachers	1070	1304	1836		
2009	**0.186**	**0.334**	**0.373**	**1.000**	
# of Teachers	1026	1199	1345	2101	
2010	**0.150**	**0.153**	**0.282**	**0.344**	**1.000**
# of Teachers	890	1019	1100	1516	2051

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 5: Correlation between EVAAS Math Results Across Years

	2006	2007	2008	2009	2010
2006	**1.000**				
# of Teachers	1886				
2007	**0.381**	**1.000**			
# of Teachers	1392	1888			
2008	**0.355**	**0.425**	**1.000**		
# of Teachers	1153	1399	1961		
2009	**0.266**	**0.277**	**0.343**	**1.000**	
# of Teachers	1008	1170	1475	1949	
2010	**0.267**	**0.314**	**0.335**	**0.357**	**1.000**
# of Teachers	844	973	1192	1434	1891

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 6: Correlation between EVAAS Reading Results Across Years

	2006	2007	2008	2009	2010
2006	**1.000**				
# of Teachers	2040				
2007	**0.270**	**1.000**			
# of Teachers	1408	1975			
2008	**0.194**	**0.310**	**1.000**		
# of Teachers	1224	1385	2046		
2009	**0.131**	**0.166**	**0.279**	**1.000**	
# of Teachers	1002	1105	1437	1947	
2010	**0.166**	**0.156**	**0.267**	**0.272**	**1.000**
# of Teachers	890	945	1205	1395	1910

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 7: Correlation between EVAAS Science Results Across Years

	2006	2007	2008	2009	2010
2006	**1.000**				
# of Teachers	1216				
2007	**0.305**	**1.000**			
# of Teachers	806	1237			
2008	**0.275**	**0.383**	**1.000**		
# of Teachers	653	844	1286		
2009	**0.120**	**0.303**	**0.439**	**1.000**	
# of Teachers	568	692	919	1300	
2010	**0.210**	**0.267**	**0.361**	**0.461**	**1.000**
# of Teachers	462	529	684	868	1237

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 8: Correlation between EVAAS Social Studies Results Across Years

	2006	2007	2008	2009	2010
2006	**1.000**				
# of Teachers	1171				
2007	**0.345**	**1.000**			
# of Teachers	836	1315			
2008	**0.282**	**0.446**	**1.000**		
# of Teachers	680	906	1335		
2009	**0.270**	**0.454**	**0.542**	**1.000**	
# of Teachers	578	709	938	1311	
2010	**0.242**	**0.311**	**0.382**	**0.427**	**1.000**
# of Teachers	486	585	732	886	1294

All correlations are at the $p<0.01$ significance level unless otherwise indicated.

Correlations between EVAAS results across subjects within a given year

As seen in tables 9-13, within a single year, the correlations between EVAAS results in the five tested subjects ranged from 0.087-0.559. These average correlations were slightly lower than the average correlations in EVAAS results within a single subject across years, although these average results increased from 0.199 in 2006 to 0.373 in 2010. The lowest correlations were between Science and Language Arts, Reading, and Math, and the highest correlations were between Science and Social Studies. Also, some of the correlations with Science were not statistically significant and are indicated on the tables below.

Table 9: Correlations Between EVAAS Results Across Subjects in 2006

	Language Arts	Math	Reading	Science	Social Studies
Language Arts	**1.000**				
# of Teachers	*2028*				
Math	**0.164**	**1.000**			
# of Teachers	*1243*	*1886*			
Reading	**0.251**	**0.336**	**1.000**		
# of Teachers	1781	1212	2040		
Science	**-0.004***	**0.221**	**0.172**	**1.000**	
# of Teachers	709	750	681	1216	
Social Studies	**0.227**	**0.136**	**0.148**	**0.336**	**1.000**
# of Teachers	815	717	773	746	1171

All correlations are at the p<0.01 significance level unless otherwise indicated.
* Not statistically significant.

Table 10: Correlations Between EVAAS Results Across Subjects in 2007

	Language Arts	Math	Reading	Science	Social Studies
Language Arts	**1.000**				
# of Teachers	*2078*				
Math	**0.226**	**1.000**			
# of Teachers	*1164*	*1888*			
Reading	**0.258**	**0.372**	**1.000**		
# of Teachers	1798	1131	1975		
Science	**0.111**	**0.193**	**0.087***	**1.000**	
# of Teachers	632	679	613	1237	
Social Studies	**0.391**	**0.274**	**0.255**	**0.510**	**1.000**
# of Teachers	881	715	852	708	1315

All correlations are at the p<0.01 significance level unless otherwise indicated.
*p<0.05

Table 11: Correlations Between EVAAS Results Across Subjects in 2008

	Language Arts	Math	Reading	Science	Social Studies
Language Arts	**1.000**				
# of Teachers	*1836*				
Math	**0.207**	**1.000**			
# of Teachers	*1174*	*1961*			
Reading	**0.269**	**0.383**	**1.000**		
# of Teachers	1613	1116	2046		
Science	**0.072***	**0.300**	**0.224**	**1.000**	
# of Teachers	645	740	599	1286	
Social Studies	**0.359**	**0.185**	**0.286**	**0.559**	**1.000**
# of Teachers	868	723	826	735	1335

All correlations are at the p<0.01 significance level unless otherwise indicated.
* Not statistically significant.

Table 12: Correlations Between EVAAS Results Across Subjects in 2009

	Language Arts	Math	Reading	Science	Social Studies
Language Arts	**1.000**				
# of Teachers	*2101*				
Math	**0.316**	**1.000**			
# of Teachers	*1118*	*1949*			
Reading	**0.311**	**0.512**	**1.000**		
# of Teachers	1743	1059	1947		
Science	**0.240**	**0.350**	**0.408**	**1.000**	
# of Teachers	611	710	566	1300	
Social Studies	**0.433**	**0.276**	**0.360**	**0.532**	**1.000**
# of Teachers	847	680	818	715	1311

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 13: Correlations Between EVAAS Results Across Subjects in 2010

	Language Arts	Math	Reading	Science	Social Studies
Language Arts	**1.000**				
# of Teachers	2051				
Math	**0.405**	**1.000**			
# of Teachers	1035	1891			
Reading	**0.363**	**0.494**	**1.000**		
# of Teachers	1709	968	1910		
Science	**0.274**	**0.306**	**0.291**	**1.000**	
# of Teachers	561	682	524	1237	
Social Studies	**0.370**	**0.356**	**0.342**	**0.530**	**1.000**
# of Teachers	805	648	784	666	1294

All correlations are at the p<0.01 significance level unless otherwise indicated.

Correlations of EVAAS and Comparative Growth Results

As seen in table 14, the correlations between the EVAAS and Comparative Growth results within a single subject and grade level ranged from 0.243-0.607, which are generally higher than the EVAAS results over time or across subjects. The correlations in the Math results are higher than the Reading results and average around 0.55. Tables 15-17 show the correlations in CG and EVAAS results within a single grade level across subjects. These correlations are also higher than the correlation in EVAAS results within a single year, with an average correlation of 0.45. Also all of the results

from these correlations were statistically significant, except for the correlation between the Social Studies Comparative Growth and EVAAS results.

Table 14: Correlation between Comparative Growth and EVAAS Results by Grade and Subject

	Math	**Reading**	**Science**	**Social Studies**
3rd Grade	**0.600**	**0.525**	-	-
# of Teachers	599	598	-	-
4th Grade	**0.537**	**0.364**	-	-
# of Teachers	513	513	-	-
5th Grade	**0.558**	**0.431**	**0.413**	-
# of Teachers	358	374	377	-
6th Grade	**0.607**	**0.533**	-	-
# of Teachers	154	175	-	-
7th Grade	**0.484**	**0.243**	-	-
# of Teachers	148	147	-	-
8th Grade	**0.505**	**0.302**	**0.381**	**0.104***
# of Teachers	150	139	130	103

All correlations are at the $p<0.01$ significance level unless otherwise indicated.
* Not statistically significant.

Table 15: Correlation between 3rd Grade Results Between Models

	Math CG	**Math EVAAS**	**Reading CG**	**Reading EVAAS**
Math CG	**1.000**			
# of Teachers	599			
Math EVAAS	**0.600**	**1.000**		
# of Teachers	599	599		
Reading CG	**0.564**	**0.476**	**1.000**	
# of Teachers	438	438	598	
Reading EVAAS	**0.406**	**0.544**	**0.525**	**1.000**
# of Teachers	438	438	598	598

All correlations are at the $p<0.01$ significance level unless otherwise indicated.

Table 16: Correlation between 4th Grade Results Between Models

	Math CG	Math EVAAS	Reading CG	Reading EVAAS
Math CG	**1.000**			
# of Teachers	513			
Math EVAAS	**0.537**	**1.000**		
# of Teachers	513	513		
Reading CG	**0.546**	**0.322**	**1.000**	
# of Teachers	330	330	513	
Reading EVAAS	**0.271**	**0.446**	**0.364**	**1.000**
# of Teachers	330	330	513	513

All correlations are at the p<0.01 significance level unless otherwise indicated.

Table 17: Correlation between 5th Grade Results Between Models

	Math CG	Math EVAAS	Reading CG	Reading EVAAS	Science CG	Science EVAAS
Math CG	**1.000**					
# of Teachers	358					
Math EVAAS	**0.558**	**1.000**				
# of Teachers	358	358				
Reading CG	**0.564**	**0.476**	**1.000**			
# of Teachers	187	187	374			
Reading EVAAS	**0.356**	**0.473**	**0.431**	**1.000**		
# of Teachers	187	187	374	374		
Science CG	**0.501**	**0.351**	**0.477**	**0.330**	**1.000**	
# of Teachers	222	222	180	180	377	
Science EVAAS	**0.345**	**0.344**	**0.274**	**0.327**	**0.413**	**1.000**
# of Teachers	222	222	180	180	377	377

All correlations are at the p<0.01 significance level unless otherwise indicated.

IMPLICATIONS

These results are in accord with the existing body of research (shown in Table 1) on the year-to-year correlation of teacher value-added results. The moderate correlations in teacher value-added results support that including student achievement data can help inform teacher evaluations, but is not the silver bullet in measuring teacher effectiveness.

The results from this case study also highlight the need for further investigation in cases where the correlations are not statistically significant. A few of the comparisons in the data set showed dramatically different effectiveness results as measured through EVAAS and Comparative Growth, which need to be fully explored. Although the correlations found in this case study are not substantially difference from the existing research on teacher effects from year-to-year, policymakers and educators must still discuss the applicability of making high stakes decisions based on measures that are somewhat volatile over time.

Further research on the consistency and stability of value-added models will help to increase the effectiveness and utility of using these models, especially when included as part of teacher evaluation. This study contributes to the research base through additional results about the correlations between subjects beyond Math and Reading, which are typically excluded from these types of evaluations.

Ultimately, making the best use of value-added results involves correlating these measures with not just other test-based results, but with observational and other pieces of data on teacher performance. Also, carefully defining the acceptable levels of performance within each of these measures would help ensure that average teachers are not penalized based on measurement error in the assessments. Finally using multiple years of value-added and student growth results would help to average teacher effects over time and hopefully improve these correlations, and ultimately increase the accuracy of conclusions drawn about teacher effectiveness.

Chapter 5: Recommendations for the Use of Value-Added Measures

This chapter examines the strengths and limitations of value-added and student growth measures in practice and explores the long-term implications for reaching a more holistic picture of teacher effectiveness. It is important to consider the ultimate goals and incentives in the use value-added measures and the utility in using these measures in defining teacher effectiveness. A comprehensive picture of teacher quality must be continually refined to ensure that the impact of a teacher on student test scores does not become representative of the complex role of teachers and schools in society. Ultimately, any trend in education reform must serve the fundamental purpose of improving teaching and learning to maximize the positive impact of schools for students.

IMPLICATIONS

Although the use of value-added measures may provide a subjective measure of teacher effectiveness, the implications of these measures and their validity and consistency over time must be fully considered. Numerous researchers have notes that getting value-added right is context dependent.[162] Both the choice of the value-added model and the ways that the measures are used need to be considered in light of the unique state or district context. Deciding on which variables to include in a value-added model and how to interpret and tie accountability measures to the results may vary greatly in different settings, and these choices have a major influence in the effectiveness of the use of these models. Also in deciding to use value-added models to inform evaluations of teacher effectiveness, a state or district is approving of the validity and

[162] Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2011). Evaluating Value-Added Models for Estimating Teacher Effects. In *Society for Research in Educational Effectiveness*. Presented at the Society for Research in Educational Effectiveness, Washington, DC.

stability of the measures over time. These measures should be carefully evaluated for consistency to ensure that value-added models are picking up a true teacher effect and not a large amount variation from random noise.

IMPLEMENTATION OF VALUE-ADDED MEASURES

In addition to the implications of these measures in validity and stability, other major considerations in implementation include the ultimate goal for the use of value-added measures and their utility as incentives. Before using any value-added measure, the goals and values of education decision makers should be made explicit to shape how a value-added model will be used.[163] Although value-added measures provide an additional source of information to inform teacher quality, states and district need to consider a comprehensive teacher evaluation model in which value-added results are simply one component.[164] While experts agree that value-added models are imperfect measures, they disagree whether those imperfections preclude value-added data from being a useful tool in evaluating teachers. One value-added expert, Dan Goldhaber, writes, "The question, however, should not be whether this is good or bad for teachers, but whether the number of incorrect classifications is acceptable given the impact on student learning."[165] In the current widely used systems of teacher evaluations, almost all teachers are rated as effective, when effectiveness truly varies along a wide range. Value-added measures can be used as a tool to help define effectiveness, but how this purpose is defined will have a

---

[163] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 5.

[164] Little, O., Goe, L., & Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/practicalGuide.pdf.

[165] Goldhaber, D. (2010). When the Stakes Are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions. Center For American Progress. Retrieved from http://www.americanprogress.org/issues/2010/12/pdf/vam.pdf.

huge impact on these measures efficiency. The definition of clear goals for the use of value-added measures will shape the ultimate impact and effectiveness of the measures for a state or district.

One of the fundamental questions about the utility of value-added models is if these measures will serve to attract and retain a more talented work force.[166] Some policymakers argue that stronger accountability measures may keep high-quality teachers from leaving the classroom and ultimately reform the framework around the profession. If the goal of these measures is to more accurately identify teacher's contribution to student learning, schools may have a valuable source of information to recognize high performing teachers, offer further development to those who are aren't getting strong results, and ultimately counsel ineffective teachers out of the profession. If value-added results are used to inform teacher practice, it may be possible to increase the mean teacher contribution to student learning.

Defining Success

Another essential question in the use of value-added models is how success is defined in educational outcomes. Although value-added models provide information about expected student growth over time, it is still possible for students to be on a low-growth trajectory and never achieve high levels of academic achievement. Many value-added models are based on normative results in which schools or teachers are defined as performing either above or below average compared with other teachers, schools, or

[166] Hull, J. (2011). Building a Better Evaluation System: Full Report. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Building-A-Better-Evaluation-System/Building-A-Better-Evaluation-System.html

statewide results.[167] These estimates of value-added have meaning only in comparison to average estimated effectiveness and do not reflect absolute level of student achievement. Ultimately measures of both proficiency and growth must inform a comprehensive understanding of student progress and achievement.

In addition to considerations to the purpose and definition of success with value-added measures, an explicit discussion of the high-stakes versus low-stakes use of these measures and their instructional relevance are essential. In high-stakes decision-making, value-added models must be held to higher standards of reliability and validity.[168] It is essential that states and districts define an acceptable level of uncertainty for these models to account for the possible misclassification of teachers in true effectiveness levels. Many states and districts are choosing to use value-added models as one source of information in informing teacher evaluation to minimize uncertainty from the results of value-added measures alone.

Instructional Relevance

Another fundamental question in the use of value-added measures is the instructional relevance of value-added results. Besides their use a measure of current teacher effectiveness, policy makers are exploring the use of these models to improve teaching. Some school districts are used value-added results as a piece in a coherent package a teacher's student achievement results with feedback on specific strengths and weaknesses in their practice.[169] Used a one portion of a larger picture, value-added results may help schools and teachers identify areas to improve effectiveness. These results may suggest which subject, grades, and groups of students the school is adding most value and

[167] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 24.
[168] Ibid, 12.
[169] Ibid, 31.

where improvement is needed.[170] Also these results can facilitate an analysis of the relationship between school inputs and school performance and can be used to create projections of school performance that can assist in planning, resource allocation, and decision-making.[171] If results from value-added models can be used to allow teachers to reflect on instruction and receive professional development, these measures will more effectively facilitate the key goals of improving instruction and student learning.[172] Ultimately states and districts must innovate in the ways that value-added results are used to strengthen the total range of feedback available to teachers and facilitate improvement in teacher practice to produce the best student achievement results.

Use of Incentives

In addition to the instructional relevance of value-added results, the use of incentives tied to these measures has a significance impact for how they are received and their utility as a tool to improve teacher effectiveness. Subtle difference in the structure of incentives can be crucial in determining their effects. Starting with a clear definition of success, incentive performance measures must align with desired outcomes. The size and structure of the consequences will also affect how the incentives operate. Incentives can be discouraging if people lack the capacity or support to reach the target that provides a reward or avoids a sanction. Incentives also need to be framed and communicated in ways that reinforce people's commitment to the goal that incentives have been put in

---

[170] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 17.
[171] Ibid.
[172] Little, O., Goe, L., & Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/practicalGuide.pdf.

place to achieve, rather than in way that erode that commitment.[173] Ultimately all incentive programs should be carefully studied to help determine which forms of incentives are successful in education and which are not.

As illustrated in the case study in the prior chapter, the stability of the results from value-added models will influence teacher responses to the incentives tied to these measures. Pay-for-performance systems based on yearly results may lead to short-term improvement since teachers will work harder for a bonus. In the long-term instability in these measures may appear more like luck, removing the incentive to change behavior.[174] The use of these measures in performance pay may also increase competition between schools and teachers, which may discourage collaboration and negatively influence school culture. The use of value-added results in incentives must be carefully evaluated to ensure that the incentive structure is a cost-effective method of increasing desired student outcomes over time.

On-going Challenges

Another important consideration in the use of value-added measures is on-going challenges with the ability of data and systems to use in generating results, the limitations of standardized tests in measuring true student knowledge, and the correlation of value-added with results from other sources of information about teacher effectiveness. Although many states and district are shifting towards value-added models for measuring student growth, many sites face limitations in the availability of statewide data systems

---

[173] National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[174] Braun, H., Chudowsky, N. and Koenig, J. *Getting Value Out of Value-Added*. Washington, D.C.: The National Academies Press, 24.

and data sets from individual schools and districts. Missing data from student and teacher mobility is a major challenge in ensuring the validity of these measures, especially in high-stakes decisions. In addition to these challenges, there is a need to develop more comprehensive methods for estimating pre- and post-measures of pupil learning in a context where states and districts lack fall-to-spring measures that are vertically scaled and reflect the full range of learning goals.[175] Ultimately value-added measures can only be as accurate and reliable as the tests on which they are based. Since the contexts of teaching are integral to the concept of teacher effectiveness, the development and use of adaptive student tests that measure a broader range of learning gains will help increase the reliability and stability of value-added results.[176] In addition to the limitations in current student assessments, another challenge in the implementation of value-added measures is in the use of multiple sources of information about teacher effectiveness to ensure validity between observational and value-added results. Multiple sources of information about teacher practice will lead to a better understanding of teacher effectiveness and better human capital decisions to lead to higher student achievement.[177]

UTILITY OF VALUE-ADDED MEASURES

In consideration of the many challenges in the implementation of value-added measures, the underlying utility of these measures must be assessed in accordance with their benefits and limitations. Adopting value-added systems in practice assume that these

---

[175] Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810, 17.
[176] Ibid, 19.
[177] Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf, 30.

measures are meaningful, reliable, and relatively stable indicators of teaching effectiveness.[178] All tools can be used incorrectly, so ensuring that a tool is used effectively is the most important to producing the best results. In Randi Weingarten's introduction to *Value-Added Measures in Education*, she states "value-added's imprecision need not be a deal breaker as long as we understand where it comes from and how to account for it when these measures are used in schools. We cannot expect any measures of teacher quality-value-added or others- to be perfect."[179] Any tool will have limitations in its use, but value-added measures can serve as one tool designed to measure student growth in ways that are meaningful to teachers. To maximize the utility of value-added measures, it is important to minimize imprecision in value-added and combine its use with other measures to provide a complete picture of teacher effectiveness.

Another challenge in the utility of value-added measures is using these results for evaluations of teacher effectiveness before they are deemed as credible by relevant stakeholders. Without a fundamental level of trust in the validity of value-added results, teachers may not use the measures in ways that can improve teaching and learning outcomes.[180] States and districts using value-added measures must also address possible gaps in understanding of how the choice of an outcome measure and teachers' own stake in the test outcome affect inferences about teacher effectiveness. The definition of accountability policies will alter teacher behavior in unintended ways and influence

---

[178] Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA.

[179] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, viii.

[180] Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge, 15.

inferences about teacher quality.[181] All of these limitations of value-added models don't necessarily mean that these measures are not useful in informing discussions of teacher effectiveness. Value-added models should be compared to the current practice in teacher evaluations and not an ideal, error free model.[182] If value-added models can be used to complement observational evaluations and other sources of information about teachers, these measures will provide additional information to draw conclusions about teacher effectiveness. It is also essential to consider the consequences for students and not just teachers in using these measures.[183] The utility of value-added information in challenging the current paradigm that all teachers are satisfactory may add value to seeing the true range of teacher effectiveness. All evaluation instruments have certain limitations, but the ultimate utility in value-added models comes from their ability to lead to desired outcomes.

DEFINING TEACHER EFFECTIVENESS

One of the additional important implications of using value-added measures in the way these methods may influence the definition of teacher effectiveness. Value-added measures may serve as one source of information about teacher quality, but they do not represent the total of expected outcomes for teachers. One the major problems with the focus on accountability is that test scores are imperfect measures of student learning that don't include important outcomes like creativity and social awareness. While the United States pushes towards a greater focus on high-stakes testing, most of the rest of the world

---

[181] Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA.
[182] Goldhaber, D. (2010). When the Stakes Are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions. Center For American Progress. Retrieved from http://www.americanprogress.org/issues/2010/12/pdf/vam.pdf.
[183] Ibid.

is focusing less on testing for standardized academic skills and more on creativity.[184] A holistic definition of teacher quality may include an initial set of qualifications to be met before entering a classroom and an on-going measure of effectiveness to evaluate a teacher's results in producing student learning.[185] In the quest to maximize teacher effectiveness, school systems need seamless transitions between pre-service, initial licensing and renewal, and evaluation. Once districts can build integrated systems for impacting education, teachers will face a common set of expectations from their initial training to a variety outcomes of their students throughout their years in the classroom.

Another issue in defining teacher effectiveness is the importance of building the professionalism of the field of teaching and giving teachers the freedom to use creativity to give students what they need. From the federal level to the individual classroom, there needs to be an essential cultural emphasis on learning, not test results, as the most desired results for students. In any accountability system and the accompanying measures for evaluation, the impact on behaviors and beliefs about teachers and education needs to be fully considered as part of the unintended consequences. In addition, the impact on non-tested teachers whose results with students aren't as easily measured has only recently entered into the discussion. Moving forward, schools and districts must reconsider their priorities in the desire to build holistic educators who can build holistically skilled students and if the current focus on test scores is achieving these results. An accountability system with measures at all levels for holistic student outcomes has the

---

[184] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 23.

[185] Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf, 46.

potential to improving teaching and learning to bring the best possible results for students.

REACHING A COMPREHENSIVE PICTURE OF EDUCATION REFORM

In order to truly reach a comprehensive picture of education reform, policy makers and educators across the nation need to promote a cultural shift in the framework about education. Instead of focusing on minimum standards of proficiency for both teachers and students, we need high standards of excellence that look forward with high goals for students and provide the support and resources to reach these outcomes. One method for achieving these goals is using targeted interventions to focus on systematic groups of teachers and students facing challenges. We need to confront the existing trends in low achievement for minorities and low-income students in a way that brings equity without expecting equality in inputs.

Ultimately the economic value of education in society stands as one essential lever in creating long-term systemic change and long-term prosperity. In order to remain at a place of international economic competitiveness, the United States must align the education system to meet a knowledge sector economy that values skills in collaboration, ingenuity, and many others beyond basic reading and math proficiency. Results on our own National Assessment of Education Progress have shown some growth in academic skills in the elementary grades, but these improvements have not materialized at the high school level.[186] Policymakers must consider if test-based accountability is serving the ultimate purpose of raising student outcomes in light of evidence to the contrary.[187]

---

[186] Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, 23.

[187] National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael

The current system may face challenges in lack of capacity and support for school improvement, but the design of accountability and evaluation systems is an essential policy tool that federal, state, and local education agencies can use to push towards desired outcomes. As highlighted in the PISA 2009 discussion of successful school policies and practices, "the quality of an education system cannot exceed the quality of its teachers and principals, since student learning is ultimately the product of what goes on in classrooms."[188] The ultimate question of how can all students receive a quality education will need to be continually re-asked and re-answered to craft the best possible educational system possible for all students.

Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[188] OECD. (2010). *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. OECD Publishing. Retrieved from http://www.oecd.org/dataoecd/11/16/48852721.pdf.

Bibliography

Amrein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. Educational Researcher. Retrieved from http://edr.sagepub.com.ezproxy.lib.utexas.edu/content/37/2/65.full.pdf+html.

Betebenner, D. W. (2008, February 8). A Primer on Student Growth Percentiles. *Colorado Department of Education*. Retrieved March 8, 2011, from http://www.cde.state.co.us/cdedocs/Research/PDF/Aprimeronstudentgrowthpercentiles.pdf

Betebenner, D. W. (2008, June 16). Presentation on the Colorado Growth Model. *Colorado Department of Education*. Retrieved from http://www.cde.state.co.us/cdedocs/Research/PDF/betebenner_norm_crit_measuresofgrowth.pdf

Braun, H. Chudowsky, N., and Koenig, J. eds. (2010). Getting Value Out of Value-Added: Report of a Workshop. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability; *National Research Council*. Washington, D.C.: The National Academies Press.

Buckley, K. & S. Marion. (2011). A Survey of Approaches Used to Evaluate Educators in Non-tested Grades and Subjects. National Center for the Improvement of Educational Assessment. Retrieved from http://colegacy.org/news/wp-content/uploads/2011/10/Summary-of-Approaches-for-non-tested-grades_7-26-11.pdf.

Campbell, J., Kyriakides, L., Muijis, D., and W. Robinson. (2004). *Assessing Teacher Effectiveness: Developing a differentiated model*. RoutledgeFalmer.

Chetty, R., Friedman, J. & J. Rockoff. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. National Bureau of Economic Research. Retrieved from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf.

Coleman, J. et al. (1966). Equality of Educational Opportunity. National Center for Education Statistics. Retrieved from http://www.eric.ed.gov/PDFS/ED012275.pdf.

Colorado Growth Model. (n.d.). Retrieved February 14, 2012, from http://www.cde.state.co.us/research/GrowthModel.html.

Corcoran, S., Jennings, J., & A. Beveridge. (2010). Teacher Effectiveness on High- and Low-Stakes Tests. In *Thirty-Second Annual APPAM Research Conference*. Presented at the Thirty-second Annual APPAM Research Conference, Boston, MA.

Corcoran, S. P. (2010). Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform. Retrieved from http://www.annenberginstitute.org/pdf/valueAddedReport.pdf

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *8*(1). Retrieved from http://epaa.asu.edu/ojs/article/view/392.

Department of Education. Race to the Top Fund. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html.

Department of Education. Teacher Incentive Fund. Retrieved from http://www2.ed.gov/programs/teacherincentive/index.html.

Earley, P., Imig, D., & N. Michelli, Eds. (2011). *Teacher Education Policy in the United States: Issues and Tensions in an Era of Evolving Expectations*. Routledge.

Fusarelli, L. (2004). The Potential Impact of the No Child Left Behind Act on Equity and Diversity in American Education. *Education Policy, 18:71*. Retrieved from http://sitemaker.umich.edu/tabbye.chavous/files/fusarelli2004.pdf.

Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. *National Comprehensive Center for Teacher Quality*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R, & A. Williams. (2005). Policymakers' Guide to Growth Models for School Accountability: How do Accountability Models Differ? The Council of Chief State School Officers, Washington, D.C. Retrieved from http://www.ccsso.org/Documents/2005/Policymakers_Guide_To_Growth_2005.pdf.

Gordon, R., Kane, T.J., & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2011). Evaluating Value-Added Models for Estimating Teacher Effects. In *Society for Research in Educational Effectiveness*. Presented at the Society for Research in Educational Effectiveness, Washington, DC.

Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press.

Harris, D., & T. Sass. (2006, April 3). Value-Added Models and the Measurement of Teacher Quality. *Institute of Education Sciences*. Retrieved from http://myweb.fsu.edu/tsass/Papers/IES%20Harris%20Sass%20EPF%20Value-added%2014.pdf

Hout, M. & S. Elliott, Eds. (2011). Incentives and Test-Based Accountability in Education. Washington, D.C.: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12521.

Hull, J. (2011). Building a Better Evaluation System: Full Report. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Building-A-Better-Evaluation-System/Building-A-Better-Evaluation-System.html

Isenberg, E. & H. Hock. (2011). Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010-2011 School Year. Final Report. Retrieved from http://ddot.dc.gov/DCPS/Files/downloads/In-the-Classroom/Design%20of%20Value-Added%20Models%20for%20DCPS%202010-2011.pdf.

Kane, T. J., Cantrell, S., Atkinson, M., Caldwell, N., Danielson, C., Ferguson, R., Gitomer, D., et al. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *National Bureau of Economic Research Working Paper Series*, *No. 14607*. Retrieved from http://www.nber.org.ezproxy.lib.utexas.edu/papers/w14607.

Little, O., Goe, L., & Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from http://www.tqsource.org/publications/practicalGuide.pdf.

Lockwood, J. R., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., and F. Martinez. (2006.) The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. The RAND Corporation. Retrieved from http://www.rand.org/pubs/reports/2009/RAND_RP1269.pdf.

McCaffrey, D. F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2004). The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness. RAND Education. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_briefs/2005/RAND_RB9050.pdf

McCaffrey, D., Sass, T., Lockwood, J.R., & K. Mihaly. (2009). The Intertemporal

Variability of Teacher Effect Estimates. *American Education Finance Association*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2009.4.4.572.

Measures of Effective Teaching Project. (2010). *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.gatesfoundation.org/highschools/Documents/met-framing-paper.pdf

Millman, J. ed. (1997). *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Corwin Press, Inc.

National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education*. Committee on Incentives and Test-Based Accountability in Public Education, Michael Hout and Stuart W. Elliot, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Newton, X., Darling-Hammond, L., Haertal, E., & E. Thomas. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Educational Policy Analysis Archives, 18* (23). Retrieved from http://epaa.asu.edu/ojs/article/view/810.

Obama, B. (July 2009). *Remarks by the President on Education*. Speech presented at the Department of Education. Washington, D.C.

OECD. (2010). *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. OECD Publishing. Retrieved from http://www.oecd.org/dataoecd/11/16/48852721.pdf.

O'Malley, K., Auty, W., Bielawski, P., Bernstein, R., Boatman, T., Deeter, T., Goldschmidt, P., Hirata, G., Hovanetz, C., Michna, G., Nedley, S., Odneal, B., Reihm, J., Stillman, L., and R. Blank. (2009). Guide to United States Department of Education Growth Model Pilot Program 2005-2008. The Council of Chief State School Officers, Washington, D.C. Retrieved from http://www.ccsso.org/Documents/2009/Guide_to_United_States_2009.pdf

Osborne, J. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=1.

Pechenone, R.L. & Wei, R.C. (2009). Review of "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences." Boulder and Tempe: *Education and the Public Interest Center & Education Policy Research Unit*. Retrieved from http://epicpolicy.org/thinktank/review-Widget-Effect.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of Value-Added Models for Estimating School Effects. *Educational Finance and Policy*, *4*(4), 492-519.

Rivkin, S.G (2007, November). *Value-Added Analysis and Education Policy*. National Center for Analysis of Longitudinal Data in Education Research (Policy Brief no. 1). Retrieved from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, *94*, 247-252.

Rosenshine, B. (1970). The Stability of Teacher Effects Upon Student Achievement. *Review of Educational Research*. Retrieved from http://rer.sagepub.com/content/40/5/647.

Rothstein, J. (2008). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w14442.pdf.

Sanders, W., Wright, P.S., Rivers, J., & J. Leandro. (2009). A Response to Criticisms of SAS EVAAS. SAS White Paper. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf.

Sanders, W. (2000). Value-Added Assessment from Student Achievement Data: Opportunities and Hurdles. *Journal of Personnel Evaluation in Education, 14(*4), 329-339.

Santos, F. & R. Gebeloff. (2012, February 24). Teacher Quality Widely Diffused, Ratings Indicate. *New York Times*. Retrieved on March 27, 2012, from http://www.nytimes.com/2012/02/25/education/teacher-quality-widely-diffused-nyc-ratings-indicate.html?_r=1&ref=robertgebeloff.

Sass, T. (2008). The Stability of Value-Added Measures of Teacher Quality and

Implications for Teacher Compensation Policy. National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/uploadedpdf/1001266_stabilityofvalue.pdf.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97-118. doi: 10.1257/jep.24.3.97

Steele, J. Hamilton, L. & B. Stecher. (2010). Incorporating Student Performance Measures into Teacher Evaluation Systems. RAND Corporation. Retrieved from http://www.rand.org/pubs/technical_reports/TR917.

Strauss, V. (2011, January 13). The Answer Sheet - New analysis Challenges Gates Study on Value-Added Measures. *Washington Post*. Retrieved January 14, 2011, from http://voices.washingtonpost.com/answer-sheet/research/new-analysis-challenges-gates-.html

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., et al. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11-36.

U.S. Department of Education. (2010). *Measuring Teacher Effectiveness Using Growth Models: A Primer* (Working Document). Race to the Top Technical Assistance Network. Washington, D.C.: U.S. Department of Education. Retrieved from http://www.ksde.org/Portals/0/Growth%20ModelsN/teacher%20growth%20models.primer.pdf

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1997). A Comparison of the Results Produced by Selected Regression and Hierarchical Linear Models in the Estimation of School and Teacher Effects. In *Annual Meeting of the AERA*. Presented at the Annual Meeting of the AERA, Chicago, IL.

Weingarten, R. (January 2010). *A New Path Forward: Four Approaches to Quality Teaching and Better Schools*. Speech presented at Education Quality for the 21st Century. Washington, D.C.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. *The New Teacher Project*. Retrieved from http://widgeteffect.org/

Wright, P. S. (2010, March 10). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf

Wright, P. S., Horn, S., & W. Sanders. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education, 11,* 57-67.

Vita

Nicole Joanne Moore was born in Houston, Texas. After completing her work at the High School for Performing and Visual Arts in Houston, Texas in 2004, she entered Austin College in Sherman, Texas. During the summer of 2007, she taught English in a slum school in Kolkata, India, which opened her eyes to the difference that education can make in long-term life outcomes. She received the degree of Bachelor of Arts in Business Administration and Music from Austin College in May 2008. During the following years, she was employed as an English as a Second Language teacher at Hickory Ridge Elementary in Memphis, Tennessee through Teach For America. In August 2010, she entered the Lyndon B. Johnson School of Public Affairs at the University of Texas at Austin.

Permanent Address: 38 Grants Lake Circle

Sugar Land, Texas 77479

This report was typed by the author.