Are State Assessments Aligned to College and Career Ready Standards More Sensitive to Ambitious Instruction in Mathematics?: Evidence from Two Large Urban School Districts

By

Brooks Alexander Rosenquist

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

May 11, 2018

Approved:

Paul A. Cobb, Ph.D.

Thomas M. Smith, Ph.D.

Jason A. Grissom, Ph.D.

Matthew G. Springer, Ph.D.

ACKNOWLEDGEMENTS

This work would not have been possible without the support and encouragement of a host of individuals, to whom I am eternally grateful. I would like to thank the taxpayers and granting agencies – specifically for National Science Foundation Grant ESI-0554535 and DRL-1119122 – that funded the multi-year research project from which this data comes, and which provided me with a tremendous on-the-job training in instrument development, data collection, interviewing, coding, … and the list goes on and on. I hope my future work and any impact it may have will add to a net positive return on investment. To the entire MIST Research Team and all of those who were part of it and supported it, I thank you. On this list, I include Jackie, who on her rounds in our office was a consistent source of encouragement.

I would like to thank those who encouraged me to pursue a Ph.D. in the first place, these people who realized it was something I might be suited for before I realized it myself. In this category I would place members of the McDonald Anderson family, the Anderson Garcia family, the Cañizares Anderson family, and the Rosenquist family. During my coursework for my master's degree at Peabody, I received crucial encouragement to take that next leap from three professors in particular: Dr. Xiu Cravens, Dr. Dale Farran, and Dr. Rogers Hall; without the council of these three scholars and mentors, I would not have attempted a doctoral degree, and my life would have been very different. While we cannot be sure what the counterfactual might be, I believe that the right choice was made. I am very grateful.

During my coursework, I benefited from the expertise and instruction of a number of terrific professors at Peabody and many supportive colleagues and fellow travelers on this journey.

I thank my children for tolerating my mental and physical absences while working towards this goal. I hope that I have served as a positive role model for persistence, if not scholarship. Most of all, I thank my patient, loving, and ever-supportive partner in life, Mónica Rosenquist.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS……iii

LIST OF TABLES……..……..……..……vi

LIST OF FIGURES……..……..……..……ix

Chapter

I. Introduction……..……..……....1

II. Review of Relevant Literature……...3

Theory and implementation of standards-based reform in mathematics…..……3
Accountability assessment: Reliability and validity issues……13
Value-added measures……..……24
This analysis……40

III. Data…...……..42

Study setting……42
Teacher observational measure……...45
Data structure……..48

IV. Methods……...51

Alternate specifications for robustness checks……52
Describing the relationship between characteristics of instruction and value-added Estimates……53

V. Results…...……..….55

Dependent variable: Teacher value-added estimates: Main model…….55
Descriptive statistics: IQA composite…….59
Analysis results: Correlations: Comparisons with MET data…….62
Analysis results: Regressions……..65
Robustness check results……70

VI. Discussion…..……....73

Review of findings and interpretation……73

Limitations……..81
Implications for policy and practice……...…89
Implications for research…….92

Appendix

A. Adoption of Common Core or "College- and Career Ready" Standards in Mathematics……95

B. Select Instructional Quality Assessment (IQA) Rubrics……...99

C. Descriptions of Main model and Robustness Models for Estimating Teacher Value-added...102

D. Details of Instability of Value-added and Classroom Observations Measures over Time and across Districts……...…….……...105

E. Regression Results from Fully-interacted Models……111

F. Results from Observed Only Classes…….…….……...114

G. Chetty, Friedman, & Rockoff Methodology……..116

H. Sensitivity Analyses: Varying the Cutoff for Class-size Exclusion……119

I. Pre-CCR & CCR Grades 6-8 Math Standards in Districts B & D:Details……123

J. Regression Approaches Addressing Measurement Error in Prior Year Test Scores……..…128

K. A Description of the Influence of Measurement Error on Estimates Using Simulated Data……...135

REFERENCES……..……..142

LIST OF TABLES

Table Page

1. Comparison of MIST participating school districts with national universe of all school districts and all urban school districts……..…….43

2. Descriptive statistics of demographic characteristics of teachers in the analytical sample, compared to district student population…….……..44

3. Descriptive statistics of certification, educational attainment, and years of experience of teachers in the analytical sample, compared to district student population…….…….……45

4. Percent agreement (kappa statistic in parentheses) for the three academic rigor rubrics of the Instructional Quality Assessment (IQA), by study year…….48

5. Change in analytical sample of teacher-year cases resulting from additional restriction of number of IQA observations and number of teacher-linked student records…...……...49

6. Change in school participation in research project, Districts B & D, Years 1-7……..…50

7. Primary estimates of teacher value added descriptive statistics, by district, by pre/post CCR assessment……...56

8. Decomposition of variance of IQA composite score, by district by test regime……...61

9. Average correlations between classroom observation scores and value-added in mathematics estimates using the MET Project methodology……...63

10. Analytical sample after omission of high-influence outliers……64

11. Linear combinations of coefficients of teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years…….66

12. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years……..67

13. Percentile rank of CCR interaction coefficient from main model, within distribution of coefficients from all value-added estimates……..72

14. Change in number of mathematical standards per grade, District B & D……..…79

A1. For states not adopting and retaining the Common Core State Standards in Mathematics, evidence on date of adoption and explicit reference to “college- and career readiness”……..……..…….97

A2. For states not adopting and retaining the Common Core State Standards in Mathematics, evidence on date of adoption and explicit reference to "college- and career readiness" (Part II)……..…….98

B1. IQA Rubric 1: Potential of the Task: Did the task have potential to engage students in rigorous thinking about challenging content?.....99

B2. IQA Rubric 2: Implementation of the Task: At what level did the teacher guide students to engage with the task in implementation?.....100

B3. IQA Rubric 3: Student Discussion Following Task: To what extent did students show their work and explain their thinking about the important mathematical content?.....101

C1. Description of specification for main value-added model (following MET Project Methodology) and 14 variations on this approach, to utilize as robustness checks…102

C2. District B: Spearman rank correlation of main value-added model, robustness models……103

C3. District D: Spearman rank correlation of main value-added model, robustness models……104

D1. Between district (and assessment) variation in correlation of value-added estimates with classroom observation measures, from the Measures of Effective Teaching (MET) study……...105

D2. Year to year variation in correlation of value-added estimates with classroom observation measures, by district, from the current study……..105

D3. Year to year variation in correlation of value-added estimates with classroom observation measures, by district, from the current study……..106

D4. By year, by district IQA correlation coefficient results when IQA is included in student-level value-added specification ……..…106

D5. Difference between within-district IQA coefficients in adjacent years……..108

D6. By year, by district IQA correlation coefficient results……108

D7. Linear combinations of coefficients of teacher value-added estimates……110

E1. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years……111

E2 Linear combinations of coefficients of teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years…..……..112

F1. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years, with value-added scores restricted only to students in classes observed by MIST researchers…….114

F2 Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. Identical teacher sample as in Table F1, but with value-added scores calculated using all students……..115

G1. Percentage of teacher-year cases with only one year of data, by district……...118

G2. Percentage of observations per teacher, by district……118

H1. Decline in teacher-year sample size as cutoff for inclusion (by number of students contributing to value-added estimate) changes……...121

I1. Pre-CCR & CCR Grade 6 Math Standards, District B……123

I2. Pre-CCR & CCR Grade 7 Math Standards, District B……124

I3. Pre-CCR & CCR Grade 8 Math Standards, District B……125

I4. Pre-CCR & CCR Grades 6-7 Math Standards, District D……..…126

I5. Pre-CCR & CCR Grade 8 Math Standards, District D……...…127

J1. Pairwise correlation coefficients from value-added estimates which address measurement error in the pre-test, across a range of potential test reliabilities…….130

J2. Alpha reliabilities of state mathematics exams in a sample of exams across districts, middle-grades, and study years……..131

J3. Fully-interacted pooled ordinary least squares (POLS) regression of teacher value-added estimates using two-step eivreg procedure……132

J4. Linear combinations of coefficients from fully-interacted pooled ordinary least squares (POLS) regression of teacher value-added estimates using two-step eivreg procedure…...134

LIST OF FIGURES

Figure Page

1. Change in average teacher value-added estimates among full-participants in the analytical dataset, over time, Districts B & D……..……..57

2. Partition of variance of teacher value-added, by district, by assessment……58

3. Change in average IQA composite score over time, Districts B & D……..…60

4. Decomposition of variance of IQA composite score, by district by test regime…….61

5. Comparison of inter-district range of observation score –value-added correlation coefficients (MET data) with intra-district range of observation score –value-added correlation coefficients (MIST data)……63

6a-c. Comparison of point estimates by time period, by district, by estimation method…….….68

7a-b. IQA-CCR Test interaction term coefficients from (a) main model value-added estimates and (b) robustness check regressions using alternate value-added estimations, District B…..70

8a-b. IQA-CCR Test interaction term coefficients from (a) main model value-added estimates and (b) robustness check regressions using alternate value-added estimations, District D….71

9a-b. Comparison of graphs of teacher observation scores from MET Study with predicted values using coefficients from regression results from District D in CCR years……...76

10. Percent of total assessment points requiring high-rigor thinking, over time, by state/district…….78

11. Number of math standards, by grade, by district; pre-CCR standards compared to CCR standards……..80

A1. Status of Common Core Standards adoption, by state/federal district, as of September 2017……..…96

D1. By year, by district IQA regression coefficient results when IQA is included in student-level value-added specification……107

D2. By year, by district IQA regression coefficient results when IQA and school fixed effects are is included in student-level value-added specification……109

E1a-c. Comparison of Point Estimates by time period, by district, by estimation method ……...113

H1a-b. Histogram distribution of number of student contributing to the value-added estimate in each teacher-year observation……120

H2. Decline in teacher-year sample size as cutoff for inclusion (by number of students contributing to value-added estimate) changes……121

H3a-f. Change in point estimates, precision of point estimates as cutoff (minimum number of students contributing to teacher-year value-added estimate) changes, by district, test regime, and estimation method……122

J1a-f. Point estimates of relationship between IQA composite and teacher value-added estimates, with 95% confidence intervals, at various reliabilities of pre-test assessment score……...134

K1a-d. Changes in estimated regression coefficients of predictor variables as percent measurement error increases……138

K2. Distribution of correlation of unobserved student effect u_{it} and the calculated residuals ε_{it} (n=1,000) at different levels of reliability of the prior student achievement variable…..139

K3. Average adjusted R-squared for a regression model with prior year as the only predictor, compared to regression models with additional covariates as predictors, at different levels of reliability of student prior achievement……140

K4. Changes in estimates of the value-added – IQA coefficient by reliability in the student prior achievement variable, by reliability of the IQA observed variable……141

CHAPTER I

INTRODUCTION

In recent years, most states have taken steps to raises their expectations for what students should know and be able to in the classroom by adopting college- and career ready standards emphasizing not only content knowledge, but particularly cognitive skills like problem-solving and communication which may have the greatest applicability for their future work and studies (Conley, 2012; Conley, et al., 2011).

These policy initiatives represent the most current iteration of *standards-based education reform*, which since the 1980s, has called for greater emphasis on problem-solving and communication skills, especially in mathematics (National Research Council, 2001; National Council of Teachers of Mathematics (NCTM) 1989; SCANS Commission, 1991). The classroom practices and interactions which advance these learning goals have been described as *ambitious instruction*, the implementation of which represents a substantial departure from practices and interactions traditionally found in U.S. math classrooms (Lampert, Beasley, Ghousseini, Kazemi, & Franke, 2010; Spillane & Thompson, 1997). In recent decades, researchers and practitioners have developed a number of resources and tools to promote and support ambitious instruction in mathematics (Cobb & Jackson, 2011).

However, while state standards for mathematical learning[1] have adapted to reflect this vision over several decades, state-adopted assessments of student learning have not kept up in their ability to promote and measure these ambitious goals for teaching and learning, often instead measuring and promoting low-level skills and knowledge (Darling-Hammond & Adamson, 2010; Herman, 2004, 2008; Schoenfeld 2007; Yuan & Le, 2012). This is problematic because these assessments are used to identify success and areas for improvement within education systems (Baker, 2005). Furthermore, the results of these tests are frequently consequential for students, teachers, and schools, given the implementation of test-based

[1] For details, see Appendix A: Adoption of Common Core or "College- and Career Ready" Standards in Mathematics

accountability policies which attach rewards and sanctions to these tests results (Elliott & Hout, 2011).

The recent adoption of career and college ready standards also coincides with the adoption of new statewide accountability assessments aligned to these standards. While the development of assessments aligned to these updated and more rigorous standards represents an opportunity for states to orient their assessment and accountability systems to measuring and promoting ambitious goals for teaching and learning, it is unclear whether these tests are in fact more likely to measure or reward ambitious instruction in the classrooms. Prior research has found that assessments vary in the degree to which they are sensitive to teachers' enactment of ambitious instructional practices in the classroom (Grossman, Cohen, Ronfeldt, & Brown, 2014; Le, Lockwood, Stecher, Hamilton, & Martinez, 2009). Because assessments and their results play a crucial role in the implementation of standards-based education reform, both research and policy can be informed by the following research questions:

RQ 1: To what extent do student growth measures (i.e., value added-scores) derived from assessments prior to the adoption of college and career-ready standards reward ambitious instruction in mathematics? Do some aspects of classroom teaching tend to be more related to these teacher value-added scores than others?

RQ 2: As states modify their assessment to ones which are purportedly aligned to more rigorous standards, do these assessments demonstrate greater sensitivity to ambitious instruction in mathematics?

CHAPTER II

REVIEW OF RELEVANT LITERATURE

Theory and Implementation of Standards-Based Reform in Mathematics

This section will address very briefly the history, influences, and implementation of the standards-based reform movement in education policy, with particular attention paid to the ways in which these kinds of reforms actually impacted school mathematics. Here, I describe some of the progress made towards establishing educational systems with curricula, instruction, teacher professional development, and assessments which are aligned to and support educational standards for learning. Finally, I describe the ways in which the standards-based education reform policies of the 1980s and 1990s evolved into the No Child Left Behind (NCLB) and most recent career- and college-ready standards movements, and how systemic reform at scale still proves to be difficult to achieve in practice.

Standards-Based Education Reform in the US

Standards-based educational reform is often described as rooted in a response to the *A Nation at Risk* report of 1983, which fomented a policy debate regarding the best ways to both raise expectations for student learning and systematically monitor student achievement (Hamilton, Stecher, & Yuan, 2012; Wixson, Dutro, & Athan, 2003). Following the report's publication and the policy push which ensued, initial efforts to raise expectations and monitor student achievement at the state- and district-level often resulted in a patchwork of ad hoc curricular and structural reforms; these attempts were largely considered unsuccessful, in part because they lacked coherence and did not clearly communicate concrete expectations for student learning (Massell, 1994). However, in some locales, the structures of standards-based reform were further along in development. In the late 1970s, a number of states began establishing structures for articulating and monitoring progress on a set of very basic learning goals through the implementation of minimum-competency examinations (Linn, 2008). Furthermore, by the mid 1908s, some vanguard states began to articulate more clear and concrete goals for student learning at each grade.

While articulating rigorous goals for what students should be expected to be able to know and do is an important first step for reform, both education researchers and policymakers in the early 1990s looked back at attempts to implement the kind of improvement called for in *A Nation at Risk* and noted that little tangible improvement occurred (National Council on Education Standards and Testing, 1992; Smith & O'Day, 1991). Largely, these authors attributed lack of improvement to a lack of systemic coherence, and offered as an alternative a vision of standards-based education reform in which key components of the educational system were aligned to achieve the goals for student learning set out in the standards documents. The formulations for systemic reform and alignment that grew out of these criticisms generally identified a few key components of standards-based education reform. These included: (1) challenging academic expectations (or "standards") for student learning ; (2) the alignment of key elements of the system (including curricular tools, professional development, instruction, assessment) to aid teachers and students to meet these new, higher standards; (3) a degree of flexibility to address local needs, and (4) structures and systems to hold students, teachers, and schools accountable to meeting these learning goals (Hamilton, Stecher, & Yuan, 2008; O'Day & Smith, 1993; Smith & O'Day, 1991).

Progress towards aligned systemic elements. As mentioned above, one of the key components in this approach to standards-based education reform was the articulation of clear but rigorous expectations for student learning, coupled with curricular tools, classroom instruction, professional development, and assessments aligned with these goals (Polikoff, 2014; Roach, Niebling, & Kurz, 2008, Smith & O'Day, 1991). The following section will briefly address these elements and describe their development and function in an aligned system of standards-based education, also drawing on the illustrative case of the State of California's efforts to implement standards-based educational reform and establish aligned standards, curriculum, and assessment.

Standards and goals for student learning in mathematics. The 1983 *Nation at Risk Report* called for the adoption of measureable and more rigorous learning standards, motivated by the observation that many high school graduates were deficient in higher order thinking skills, and suggesting that "schools may emphasize such rudiments as reading and computation at the expense of other essential skills such as comprehension, analysis, solving problems, and drawing

conclusions" (p. 116). Educational research supports the view that schools in the U.S. have typically taught mathematics with an emphasis on lower-order thinking skills including memorization and the execution of mathematical procedures and algorithms (Spillane & Zeuli, 1999; Hiebert et al., 2005; Weiss, Pasley, Smith, Banilower, & Heck, 2003). Advocates of mathematics education reform in the U.S. have cited a number of rationales to motivate their calls for change (e.g. (National Commission on Excellence in Education, 1983; National Council of Teachers of Mathematics, 1989, 2000, 2014; National Research Council, 2001, 2012; SCANS Commission, 1991). For example, these advocates have cited the relatively poor performance of the U.S. students in international comparisons of educational achievement, especially compared other developed countries and countries in East Asia. Relatedly, student achievement in academic disciplines related to the fields of science, technology, engineering, and mathematics (STEM) has been posited as key to the future economic, industrial, technical, and innovative prowess of the United States, which has often been portrayed as eroding relative to other countries since the 1980s. Even when setting aside considerations of economic competitiveness and global competition, advocates for the reform of mathematics education have pointed to evidence of the transformation of the US economy from an industrial-based to serviced-based "knowledge economy," along with the increasing sophistication of technology and automation. As a result of these changes, an increasing proportion of the economy's jobs will call for nonroutine analysis, collaboration, and problem solving, with less demand for routine and manual skills (Murnane & Levy, 1996). If the U.S. educational system does not inculcate these skills in its students – the thinking goes – there is likely to be a troubling mismatch between the skills present in the labor force and those most in demand in the labor market (Darling-Hammond & Adamson, 2010; Murnane & Levy, 1996).

In the years immediately preceding publication of the *Nation at Risk* report, much of the policy response was at the state- and local-levels. In one of the earliest example of state-level response attempting system-level reform, the California Department of Education published curriculum frameworks for mathematics and language arts in 1985 and 1987, respectively, with the view that establishing a consensus on explicit and concrete goals for student learning would provide an important first step for orienting system resources and efforts in support of these goals (Carlos & Kirst, 1997; Wixson et al., 2003). This curricular framework – described as

requiring more intellectually sophisticated instruction, more engaging work for students, and an emphasis on conceptual understanding (Cohen & Hill, 1998) – became a model for other states developing their own mathematics standards (Wixson et al., 2003). In 1989, the National Council of Teachers of Mathematics (NCTM) published their *Curriculum and Evaluation Standards for School Mathematics* which built on the California framework and further emphasized the importance of conceptual understanding in the service of higher order cognitive skills, along with the student-centered, inquiry-based, hands-on, and more active approach to learning thought to best cultivate these kinds of skills. These two documents served as influential models for the remaining states which came to formulate and adopt content standards in the ensuing years, with the influence of this approach solidifying when, in 1994, the reauthorization of the Elementary and Secondary Education Act made receipt of Title I contingent on the development and implementation of content standards and aligned assessments (Wixson et al., 2003) and the U.S. Department of Education began awarding grants for states to develop their own standards (Hamilton et al, 2012). As practitioners and policy makers were confronted with evidence of within-state disparities in achievement from student of different communities, this effort to articulate explicit and uniform standards for learning was motivated to a large extent by a desire to narrow these achievement gaps through equalizing students' opportunities to learn (OTL) within a given state (Elmore & Fuhrman, 1995).

After initial adoption, state standards for mathematics education underwent periodic revision, as did those published by the NCTM (Finn, Julian, & Petrilli, 2006; Finn, Petrilli, & Vanourek, 1998; NCTM, 2000). Even as the No Child Left Behind (NCLB) Act of 2001 applied standards-based accountability logic to the nation as a whole, individual states were still allowed to dictate and define their own curricular standards. During this period, states were found to vary widely in both the specificity and rigor of their standards (Finn et al., 1998, 2006) as well as in the level of knowledge students which were required to demonstrate in end-of-the-year tests to qualify as "proficient" (National Center for Education Statistics, 2007). Some analyses have suggested that the accountability mechanisms put in place by NCLB provided perverse incentives for states to maintain low standards (Balfanz, Legters, West, & Weber, 2007; Koretz, 2008). In 2007, the Council of Chief State School Officers (CCSSO) met to discuss the formulation of a single set of education standards in mathematics and English language arts in

order to provide an alternative to the uneven patchwork of educational standards articulated at the state level (Conley, 2014). These would extend previous state-level efforts to emphasize higher order thinking and provide greater equity in students' opportunity to learn. At the same time, this effort also stressed that these updated and uniform standards should (1) be "internationally benchmarked" (i.e. they should reflect the goals and processes demonstrated by other developed nations with high-performing education systems), and they should also (2) emphasize "carrier and college readiness" (i.e. they should reflect the knowledge and skills most important for success in post-secondary education and the workplace) (Conley, 2014, p. 2). Although the mathematics and English language arts standards produced through the efforts of the CCSSO have generally been more rigorous than many of the state-level standards which proceeded them (Porter, McMaken, Hwang, & Yang, 2011) only forty-two states have adopted these standards,[2] with a handful of these states later repealing adoption (Bidwell, 2014). Nevertheless, these most recent formulations of educational standards continue to be influential, with many states writing and adopting career and college ready standards which are very similar in rigor and content to the Common Core State Standards (Conley et al., 2011).

Aligned curricular tools, instruction, and teacher professional development. While the theory of change of standards-based school reform asserts that the articulation of these more explicit and challenging goals for student learning fulfills an important systemic purpose, these goals are likely to be attained only with the alignment and support of a number of additional important elements of the educational system. In this theory of change, curricular tools, a vision of high quality instruction, teacher professional development, and student assessments aligned to and supporting these student learning goals are necessary to engender coherent systemic change. This section will address the first three elements, with assessment addressed later in its own section.

Given the ways in which curricular materials profoundly influence teachers' classroom instruction (Remillard, 2005; Tarr et al., 2008), teachers will need to have available curricular tools which are aligned to and support these ambitious learning goals. Indeed, development of these new materials required its own specific systemic effort, because traditional textbook

[2] For details, see Appendix A: Adoption of Common Core or "College- and Career Ready" standards in mathematics.

publishers were not always quick to adapt to these new standards. For example, after the adoption of the more rigorous mathematics curriculum framework, California's Department of Education initially rejected the majority of textbooks as not sufficiently supporting the new learning standards and began encouraging the independent development of modules or "replacement units" to supplement more traditional textbooks (Cohen & Hill, 1998, p.3). However, as the adoption of these kinds of more challenging goals for student learning spread to a number of states, a broader approach was needed. In 1989, the National Science Foundation (NSF) began funding the development and support of curricular materials aligned with mathematics standards like those described by the NCTM Standards, spending an estimated $93 million in the 18 years which followed (National Research Council, 2004b).

Since the writing of the NCTM standards in 1989, there has also been a considerable research and development of instructional practices which promote these more ambitious goals for student learning (Franke, Kazemi, & Battey, 2007; Kirkpatrick, Martin, & Schifter, 2003. In order to address the most recent wave of college- and career-ready standards for student learning, NCTM drew upon much of this research to update its published guidelines for instruction in mathematics (2014). This vision is of what constitutes high quality instruction in mathematics is often referred to as *ambitious teaching* (Lampert, et al., 2010). In instruction congruent with this vision, teachers support students to solve cognitively-demanding tasks (Stein, Smith, Henningsen, & Silver, 2000), press students to provide evidence for their reasoning and to make connections between their own and their peers' solutions (McClain, 2002), and orchestrate whole-class discussions to develop student thinking and build a shared understanding of mathematical concepts (Franke et al., 2007; Stein, Engle, Smith, & Hughes, 2008). Instructional practices of this type contrast sharply with typical teaching in most U.S. classrooms and require teachers to anticipate and respond to students' thinking (Kazemi, Franke, & Lampert, 2009). This vision of ambitious teaching in mathematics is also reflected in the development of a number of discipline-specific classroom observational instruments designed to describe and measure this kind of instruction (Borman, 2005; Gallimore, & Hiebert, 2000), including the Instructional Quality (IQA) Assessment instrument (Matsumura et al, 2006) and the Mathematical Quality of Instruction (MQI) instrument (Hill et al, 2008).

However, the kind of classroom instruction implicated by these standards represents a dramatic change compared to the traditional teaching of mathematics that most teachers have

been practicing or had experienced as students. Some researchers have suggested that the required change in teacher practice entails reorganization rather than only the elaboration or extension of current practices (Kazemi, Franke, & Lampert, 2009), requiring substantial investments in teacher supports over an extended period of time (Darling-Hammond, Wei, & Orphanos, 2009). As such, the successful implementation of standards-based instruction in mathematics has been said to require a reconceptualization of the knowledge, skills, roles, and dispositions teachers need to be effective in the classroom (Heck, Banilower, Weiss, & Rosenberg, 2008).

Relatively early on in the standards-based education reform movement it became clear that, given the substantial changes in teaching and learning called for with these newest standards, teacher professional development and other supports would be crucial to successful implementation of these curricula (Cohen & Hill, 1998; Goertz, Floden, & O'Day, 1995; Smith & O'Day, 1990). Teacher professional development is typically included along with standards, curriculum, instruction, and assessment as an essential element of an aligned standards-based educational system (Fishman, Marx, Best & Tal, 2003; Massell, 2000). However, quantifying the degree to which teacher professional development is aligned to mathematical standards is not as straightforward as measuring the alignment between many of the other elements, as described in alignment studies of curriculum, instruction, and assessments (e.g. Martone & Sireci, 2009; Polikoff, Porter, & Smithson, 2011; Porter, 2006). However, research in mathematics education has begun to make substantial progress in outlining effective professional development practices (Borko et al., 2009; Elliott et al., 2009; Ball, Sleep, Boerst, & Bass, 2009; Kazemi & Hubbard, 2008; Lampert, Beasley, Ghousseini, Kazemi, & Franke, 2010), including identifying a small set of concrete, teachable, and high-leverage instructional practices which support these more ambitious goals for student learning (Ball et al., 2009; Grossman et al., 2009; Lampert et al., 2010; Lampert & Graziani, 2009). Furthermore, both researchers and practitioners have begun to recognize the role which job-embedded forms of professional development – such as teachers' professional learning communities (Horn & Little, 2010; McLaughlin & Talbert, 2006) and instructional coaching – can play in developing and sustaining these practices (Miles, Odden, Fermanich, & Archibald, 2004; The New Teacher Project, 2015).

Assessments. Assessment has always played a key role in frameworks for standards-based educational systems (Hamilton, Stecher, & Yuan, 2008; O'Day & Smith, 1993; Smith & O'Day, 1991). In these frameworks, assessments serve multiple functions. Yearly results from testing provide a means for measuring systemic progress and guiding the improvement of learning, consistent with an evidence-based and scientific approach to research and development which employs iterative cycles of experimentation, data analysis, and modification (Baker, 2005). The assumption is that teachers and others involved in education policy implementation will respond to data, incentives, and sanctions to align systemic activities and structures in ways that effectively realize the goals and outcomes articulated by the system (Polikoff, 2014; Roach et al., 2008).

Following the establishment of the curriculum frameworks, the state of California designed in 1991 a standards-based assessment called the California Learning Assessment System (CLAS) for reading, writing, and mathematics. The CLAS is accurately characterized as a series of *performance assessments,* which is to say that an examinee must either construct or supply an answer, produce a product, or perform an activity (Madaus & O'Dwyer, 1999; Stecher, 2010). Performance assessments are often contrasted with multiple-choice tests, with performance assessments characterized as more authentic (i.e. presenting a more realistic assessment of the application of skills and knowledge) and more likely to measure schematic knowledge (knowing why) or strategic knowledge (knowing when, where and how to apply skills and knowledge); by contrast, multiple-choice items are described as less authentic and more likely to measure declarative knowledge (knowing that) and procedural knowledge (knowing how) (Shavelson, Ruiz-Primo, & Wiley, 2005; Stecher, 2010). The CLAS tests were first administered in 1993 to students in grades 4, 8 and 10 with the purpose of gauging student achievement at both the individual- and school-levels. In math, these assessments not only asked student to show how they arrived at their answers, but also incorporated a portfolio of student work to factor into the assessments measures of student achievement (Stecher, 2010).

California's move towards the use of performance and portfolio assessments was largely in line with reform thinking on assessment, such as the National Commission on Testing and Public Policy (1990) which advocated for the replacement of multiple-choice tests with performance assessments. The California assessments were similar to a number of other

performance and portfolio assessments integrated into the standards-based reform efforts in a number of other states in the early 1990s, including Kentucky, Maryland, Vermont, and Washington (Stecher, 2010). Consistent with research and theory which suggests that teachers will adjust the form, content, and rigor of instruction in response to changes in student assessments (see reviews in Cheng & Curtis, 2004; Herman, 2008), there is some evidence that in some of these states, these more ambitious assessments were associated with a number of changes in teacher instruction which might be interpreted as following the intent of the reform initiatives (for a review, see Stecher, 2002, 2010). For example, with teachers report being influenced by the assessment to spend more time on problem-solving, communication, and group work (Kortez, Barron, Mitchell, & Stecher, 1996).

However, in some places, these tests were controversial and did not meet with wide support from all stakeholders. Not all teachers agreed with the learning goals represented by the new standards and new assessment (Cohen & Hill, 1998). Scores from the performance and portfolio assessments were criticized by some as being excessively unreliable because of problems with interrater reliability, lack of standardization of student portfolios, and because these assessments typically consisted of a small number of relatively time-consuming performance items (Dunbar, Koretz, & Hoover, 1991; Koretz, Stecher, Klein, & McCaffrey, 1994: Shavelson, Baxter, & Gao, 1994). In California, concerns about falling NAEP scores, reliability, cost, and the secrecy of test items, together with turnover in state leadership, resulted in the CLAS being discontinued after only two years being administered (Carlos & Kirst, 1997; Marsh & Odden, 1991; Stecher, 2010) However, other similar state-level attempts at more sophisticated, standards-based assessments were implemented for a longer period, although these assessments were eventually discontinued for the purposes of statewide accountability in light of the testing and reporting requirements of NCLB (Stecher, 2010)

Shift toward test-based accountability. Guthrie and Springer (2004) characterized the era of education reform during the period from 1990 to 2000 as largely emulating models of systemic reform, especially the framework formulated by Smith and O'Day (1991). They described the post-NCLB period which followed as characterized by accountability driven by outcome measures. In practice, the seeds of outcome-based accountability were planted during earlier periods of reform. For example, holding teachers and schools accountable was part of the

early systemic improvement framework put forward by Smith and O'Day[3]. Furthermore, statewide testing systems had been established even earlier, holding students accountable for their own outcomes with minimum competency tests in the 1970s (Shepard, 2008), and the implementation of exams for promotion and graduation in some states in the decade preceding the passage of NCLB (Bishop, 1998; Goertz & Duffy, 2001, Harris & Herrington, 2006). However, during this period, a number of states also experimented with models for school-level assessment-based accountability, with sanctions ranging from public hearings to school takeovers, reconstitution, and eventual closing (Goertz & Duffy, 2001).

In what Guthrie and Springer (2004) characterized as the "third wave" of education reform, this model for school-based accountability was integrated into national education policy with the passage of the No Child Left Behind Act of 2001 (NCLB). Along with a focus on assessment-based school-level accountability consequences, NCLB also required the reporting of scores disaggregated by a number of student subgroups historically considered at-risk for lower achievement, such as English learners, students receiving special education services, and some racial and ethnic minorities. NCLB would require testing at a much larger scale than previous reforms (approaching every student, every year testing), with consequences attached to both individual- and subgroup-level performance. These policies would require tests which could provide more precise estimates of student achievement with smaller margins of error and higher reliability. Increased reliability and precision would require greater numbers of items per test; this requirement, combined with an increase in the scope of testing with more students tested every year, resulted in a strong bias away from more expensive and time consuming performance items towards a preference for multiple choice test items. (Shepard, 2008; Stecher, 2010). Perhaps it is not surprising then that most of the NCLB-era assessments overwhelmingly lack test items of high cognitive demand (Yuan & Le, 2012).

[3] E.g. "States must construct and administer high quality assessment instruments on a regular basis to monitor progress toward achievement goals for accountability purposes and to stimulate and support superior instruction." (Smith & O'Day, 1991, p. 252)

Accountability Assessment: Reliability and Validity Issues

Assessment Score Reliability

Reliability has to be taken into account when analyzing scores from student assessment, and becomes a crucial issue to attend to in the accountability context, to determine if a given assessment is suitably reliable to draw inferences about the organizational or analytical unit in question, be they schools, teachers, or students. For example, the National Assessment of Educational Progress (NAEP), in order to reduce time-taking burden on students, utilizes a matrix sampling scheme where each student is presented with a subset of relevant items, allowing for inferences to be made about students' skills and knowledge at a district or state level, but not allowing for accurate and reliable inferences at the student level (Cheong, Fotiu, & Randenbush, 2001). In addition, accountability test scores for English language learnings are less reliable than those for non-English language learners, to the extent that inferences on performance of this subgroup at the school level may not be made with adequate confidence, especially in schools with relatively low numbers of these students (Abedi, 2004). Furthermore, attention to the reliability and measurement error may in some circumstances be important in the implementation of assessment-based accountability within a school setting, as when principles are called upon to make high-stakes decisions about teachers or students based on tests scores, especially when making inferences about student proficiency when students are very close to a test score cutoff (Means, Chen, DeBarger, & Padilla, 2011).

Central to the issues of certainty around scores (i.e. their reliability) is the fact that student assessments measure student learning, but the scores produced by these tests also include some measurement error (Boyd, Lankford, Loeb, & Wyckoff, 2013). Classical test theory defines observed test scores as the sum of two components: a "true score" and a degree of measurement error. In this framework, the true score is operationalized as the expected value of a test taker's observed scores over several replicated measurements. Deviations of the observed test score from the theoretical true score are conceptualized as measurement error. Given the variation in observed scores over repeated measurements taken from one or more test takers, classical test theory can be used to estimate the proportion of error in each measurement and also extrapolate these findings to a distribution of scores, estimating the proportion of variation in the

sample which is due to measurement error and the proportion of variation which reflects "true variation" in the construct being measured. Put differently, classical test theory's framework for conceptualizing true variance and error variance allows us to describe and compare the "signal" in a distribution of test score (the proportion of total variance which is composed of variation of true scores) with the "noise" in the same distribution (the proportion of total variance which is corresponds to measurement error).

In contrast to classical test theory, generalizability theory (Brennan, 2001; Cronbach, Rajaratnam, & Gleser, 1963) provides a framework for identifying and estimating the unique contribution of a number of potential sources of measurement error. Generalizability theory describes an analog of classical test theory's "true score", the *universe score*, which is conceived of as the theoretical and unobserved average value of an individual's scores over the entire universe of possible acceptable alternative measures for the construct being measured (Haertel, 2006). In a way which continues the analogy with classical test theory, all other sources of variation which cause the observed score to differ from the universe score can be thought of as contributing to the measurement error in the observed scores. Generalizability theory departs from classical test theory in that it attempts to estimate the amount of error variance emanating from a few key characteristics (or *facets*) of the measurement: the test instrument, the test-taker, and the testing occasion.

The notion of *universe score* provides an insight into one important source of assessment measurement error: *item sampling error*. Because a measurable construct such as "eighth grade mathematics content" could consist of a large number of topics (e.g. inequalities, linear equations, operations on polynomials, etc.) at a number of different levels of cognitive demand (memorization, applying algorithmic procedures, communicating understanding, proving, etc.), it is likely that a given assessment will only measure performance on a subsample of this entire body of knowledge and skills. For that reason, some of the difference between the observed score and the unobserved universe score can be attributed to item sampling error. However, Item sampling error is only one variety of test-specific error. Other sources of test-specific measurement error may result from question and answer format and clarity of the test directions and questions (Kiplinger, 2008).

Some of these sources of error can be traced back to idiosyncrasies of individual test takers or the testing occasion. Test taker-specific sources of error include either persistent or transitory characteristics of the individual student (such as test anxiety or level of fatigue) that cause the student to respond to test items in a way which underestimates or overestimates the universe score (Kiplinger, 2008). Examples include idiosyncrasies in test administration, room temperature, crowding, distracting peers (Kiplinger, 2008), or even a barking dog in the playground outside the classroom (Kane & Staiger, 2002; Papay, 2011). In addition to sources of error emanating from the three individual facets, generalizability theory also identifies sources of potential and quantifiable error variance which result from the interaction of one or more of these facets. For example, assessment items in reading comprehension or writing may call on students to draw upon background knowledge, such that the student without this background knowledge would tend to score lower on these items (Gebril, 2009). In the framework of generalizability theory, this would be considered an item-by-person source of error.

Generalizability theory provides a broad and flexible framework for inquiry into and quantification of numerous potential sources of measurement error which may detract from the reliability of any given set of scores from an assessment for student learning. However, standardized tests developers typically report only the statistic derived from dividing a single test into parallel parts and estimating the proportion of variance in scores which is shared by these subtests. However, this kind of reliability – split test reliability – does not account for other sources of measurement error and is likely to understate all sources of measurement error variance (Boyd, Lankford, Loeb, & Wyckoff, 2013; Feldt & Brennan, 1989). For example, Boyd and colleagues (2013) found that the while test designers reported only 5 to 10 percent error variance in the math assessments used for accountability in the state of New York, more sophisticated methods which consider additional sources of error estimated the proportion of measurement error variance to be approximately 16 to 20 percent of total score variance.

Assessment Score Validity

Along with reliability, the validity of assessments and their scores need to be considered at the time of interpretation and analysis. The *Standards for Educational and Psychological Testing* (American Educational Research Association/American Psychological Association

/National Council on Measurement in Education [AERA/APA/NCME], 2014) clarify that validity is not most appropriately conceived of as a characterization of a test *per se*, but instead as a characterization of the interpretations or inferences made from test scores and entailed by proposed uses of the tests. The *Standards* also describes the process test validation as accumulating relevant types of evidence to provide a sound scientific bases for proposed interpretation of a measure's scores. The *Standards* go on to identify and describe different types of validity evidence, including evidence based on the content of the tests, as well as evidence based on the relationships observed between test scores and other variables. Because student assessments are used to make inferences about both student learning and educational effectiveness, I will address the evidence for validity for both purposes here separately.

Validity of assessments scores as measures of learning. One important dimension of the validity of a test score's meaning comes from the evidence the test content provides for *construct validity*: the degree to which a test measures what it purports to measure (Cronbach & Meehl, 1955). There are two major threats to construct validity: *construct underrepresentation*, and *construct irrelevant variance* (Messick, 1990). Whereas construct underrepresentation refers to when a test is too narrow, failing to include important dimensions of the construct represented, construct irrelevant variance refers to variation in test scores that are a function of real and systematic differences in a construct other than that which the assessment intends to measure.

One method of addressing an assessment's construct representation is through evaluating its alignment to and coverage of the body of knowledge and skills it purports to measure (Linn, 2008). The most widely used approaches to measuring alignment and coverage take into account the degree to which assessments align with both (1) the range and knowledge specified in the educational standards, as well as (2) the level of cognitive complexity (Porter, 2006). A number of studies which have looked at alignment of mathematics standards and assessment on these two dimensions for have frequently found that assessments do not adhere closely to state standards in terms of topics covered, and that the topics which are tested were usually assessed at lower levels of cognitive demand than are called for by the state standards (Resnick, Rothman, Slattery, & Vranek, 2003; Polikoff, Porter, & Smithson, 2011; Webb, 1999). For example, in one of the most recent alignment studies of a sample of NCLB-era assessments (Polikoff et al., 2011), researchers found that almost one-third of the topics in the grade-level math standards are not

reflected in the state assessments, with almost one-third of the topics covered by items on the state assessments not reflected in the grade-level state standards. Some research has attributed assessments' lack of representativeness of the larger academic domain they are purported to measure to test developers omitting items with low item-total correlation or very high or very low difficulty indices (i.e. items that are answered correctly by a very large or small proportion of the intended testing population) (Polikoff 2010; Popham, 1999; Wiliam, 2007).

Assessments used for accountability purposes have also been investigated for evidence of *concurrent validity*, or the extent that variation in these measures of student learning corresponds to variation in other measures of student learning. One consistent threat to concurrent validity of accountability assessment scores have been found in studies which look at populations with significant changes in average scores on accountability assessments and then attempt to validate these scores through similar concurrent changes in lower stakes test scores for the same population (Koretz, 2005). A number of studies have found evidence that aggregate gains in scores from assessments used for accountability purposes are not supported by concurrent gains in similar low-stakes tests administered to the same population (see reviews in Holcombe, Jennings, & Koretz, 2013; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz, 2005). This phenomenon, which suggest that test gains on accountability tests are not reflected in other lower-stakes assessments or "audit test," is often referred to as *test inflation* (Koretz, 2005). Test inflation may be the result of teachers' responses to the accountability tests, including changing instruction to include greater focus on tested subjects and standards (Jennings & Bearak, 2014) or greater instructional focus on test-specific knowledge and skills, which raises scores but does not generate transferable knowledge and skills (Klein, Hamilton, McCaffrey, & Stecher, 2000).

The validity of assessment scores as a measure of effective of instruction. In the previous section, the validity of student test scores as a measure of student learning is evaluated with two categories of evidence: (1) the degree to which the content of a test reflects the construct measured, and (2) the correlations these scores exhibit or fail to exhibit with other variables (particularly, other scores from other tests). However, when considering interpreting student scores or gains in those scores as a measure of the efficacy of instruction, it becomes vital to look deeper into the reliability of these scores and change in these scores at the teacher level.

Reliability is often, but not always, described as a precondition for validity (AERA/APA/NCME, 2014; Haertel, 2006). Even as a given assessment score is a noisy and imperfect measure of student learning, it is an even noisier and less perfect measure of the effect of teaching. The framework of generalizability theory of measurement, discussed above, addresses multiple sources of error that may make it difficult to isolate the signal of true student learning from the noise introduced by factors specific to the test, the student, the testing occasion, or any interaction between those three facets of the measurement. However, even if we were to isolate an error-free estimation of true student learning from the other sources of error variation present in the scores, the teacher contributions to this learning are only one small part of that variation in the true student learning score. Generally, the knowledge and skills acquired by a student at any given time are theorized to be a function of all prior learning experience both in and out of the school environment, along with individual characteristics, including those which change over time and those which do not (Popham, 1999; Guarino, Reckase, & Wooldridge, 2015). All of the factors that contribute to a student's learning of over time will potentially contribute that student's performance on a given test. Similarly, between-student differences in the factors will contribute to the variation between students in the distribution of scores on any given assessment. It is common for test items to be more sensitive to – and therefore better measures of – differences in out of school learning or individual differences in intellect and aptitude than they are sensitive to differences in students' in-school learning (Popham, 1999)[4]. As a result, the amount of variation in test scores which provides a signal for quality teaching is relatively low compared to the amount of variation that represents "noise" from other inputs. One analysis suggests that the vast majority of variance in student test scores

[4] Popham (1999) explained that standardized test designers seek test items with high degrees of discrimination, i.e. test items which are answered correctly by close to half of the test taking population, as opposed to tests which are answered correctly by large proportions of students or answered incorrectly by large proportions of students. Popham argued that these kinds of items tend to be more sensitive to (and therefore, are better measures of) either (a) out of school learning or (b) "native intellectual skills that are not readily modifiable in school" (p. 13), and are relatively insensitive to – and therefore poor measures of – differences in the efficacy or quality of school environments. The exact nature of the "native intellectual skill" or "in-born intellectuability" (p.13) two which Popham referred is still not well understood and contested by researchers and theorists, who point to both inter-individual correlations of tests of different kinds of mental ability (verbal, spatial, symbolic, numeric) as well as the proportion of variation estimated to be attributable to genetic differences between individuals (estimated to be 0.50 in Neisser et al. 1996). At the same time, other research has pointed to some malleability of scores on tests of intelligence as the results of schooling or other interventions, interactions with social-class, as well as evidence for discrete *crystalized* and *fluid* components of general intelligence – see Nisbett et al. (2012) for a review.

(about 60%) is explained by individual and family background characteristics, with teacher-level factors only accounting for about 8.5% of the variation, and class-level factors accounting for another 4% (Goldhaber, Brewer, & Anderson, 1999). If we are interpreting growth in student test score as a measure of teacher effectiveness, then this measure consists of roughly 10 percent signal and 90 percent noise. By contrast, when we are interpreting student test scores as a measure of student learning, recent analysis suggests that this measure consists of roughly 80 to 85 percent signal and 10 to 15 percent noise (Boyd, Lankford, Loeb, & Wyckoff, 2013). That one test might serve one purpose reasonably well (i.e. measure student learning) while serving another purpose rather poorly (i.e. measuring the effectiveness of teaching and instruction) highlights an important point of consensus in the research community: most educational tests serve one measurement purpose better than another, and there is a tension or tradeoff between the number of purpose a test can serve and its effectiveness at serving those purposes (AERA/APA/NCME, 2014).

Having established that students' aggregate tests score growth is a considerably noisy measure of teachers' contributions to learning, there are two primary approaches to addressing this issue. One approach would be to use results from tests designed to measure student learning and use sophisticated statistical methods to attempt to isolate teachers' contributions to student learning gains: this approach – referred to as *value-added modeling* and discussed at greater length in the following section – is the method currently utilized in much of the recent research which tries to quantify individual teacher effectiveness using measures of student learning. Another approach would be to construct and develop tests which are more sensitive to instruction, that is, with a greater proportion of variation in scores reflecting the quantity and quality of the classroom instruction in which the student has participated (Wiliam, 2007). While there has been some conceptual and empirical work to operationalize and develop these kinds of assessments (e.g. Ruiz-Primo et al., 2012), their formats, development, and properties would be very different from assessment of student achievement currently used for accountability purposes. For example, an eighth-grade test consisting of items which most students would answer incorrectly at the beginning of the eighth-grade year but which most students would answer correctly at the end of the year has been described as one with high instructional sensitivity (Polikoff, 2010; Wiliam, 2007). However, constructing tests with these qualities

would be difficult and counter to a number of practices currently used to develop norm-referenced tests. The distribution of student achievement for adjacent grades, as currently measured by a number of nationally normed achievement tests, has considerable overlap (Hill, Bloom, Black, & Lipsey, 2008; Wiliam, 2007).[5] Additionally the omission of test items with low item-total correlation or very high or very low difficulty indices (mentioned above) are test development practices which might also contribute to a lack of instructional sensitivity (Polikoff 2010; Popham, 1999; Wiliam, 2007). Furthermore, most definitions of the instructional sensitivity of tests dictate that test score variation should be sensitive to both the quantity and quality of classroom instruction (Ruiz-Primo et al., 2012), which might require not only arriving at a defensible or agreed upon definition of instructional quality, but which might possibly entail collecting and incorporating more explicit or direct measures of the quality of instruction quality of instruction (Polikoff, 2010).

Questions of reliability and instructional sensitivity aside, it is also important to consider the validity of student gain scores as measures of effective instruction by looking more deeply at the content of the tests themselves. Test scores cannot currently assess many of the non-cognitive student outcomes which we would hope that quality teaching would engender, such as student curiosity, interest, and motivation in academic studies (Koretz, 2002). That being said, it is not clear that student assessments used for accountability purposes are adequate measures of even the set of cognitive outcomes of schooling they are supposed to measure, i.e. the body of knowledge and skills described by the grade level content standards. Classroom teachers are charged with addressing and supporting the learning of a defined body of knowledge and skills described in the state standards for a particular grade, and the previous discussion of the validity of these tests as measures of student learning discusses the mismatch empirical studies have detected between the depth and breadth of state assessments and the depth and breadth of the standards which they are intended to reflect (e.g. Polikoff, Porter, & Smithson, 2011).

Finally, evidence for the validity of interpreting student test score gains as valid measures of teacher contributions of learning can be marshalled through investigating the way in which

[5] For example, Hill and colleagues (2008) drew from a number of nationally normed achievement tests and estimated that one year of eighth grade instruction on average advances students 0.32 standard deviations on the seventh grade scale. Put differently, on average, a student scoring at the 50th percentile on the grade 7 test at the end of grade 7 might, after one year of instruction at the eighth grade level, be expected to score at the 63rd percentile on the grade 7 test.

student test score gains correlate with other variables relevant to the context and its processes and outcomes. Most recent and rigorous analysis which seeks to investigate the relationship between instructional quality and other important educational constructs has not, in general, used average unadjusted test score gains as a proxy for instructional quality, chiefly because of their lack of reliability for this purpose and because these unadjusted test score gains can be influence by a number of confounding factors. Research of this kind has typically analyzed variables of interest vis-à-vis teacher-level value-added measures. This literature is reviewed later in this dissertation, addressing the degree to which other variables of interest lend support for or against the interpretation of teacher value-added estimates as measures of teacher effectiveness or instructional quality.

Consequential validity and systemic validities. In providing validity evidence for their measures, test developers are generally expected to articulate clearly their test's appropriate intended uses, interpretations, and testing population (AERA/APA/NCME, 2014). However, some scholars suggest more is demanded, maintaining that some of the *consequences* of test use are also squarely within the traditional conceptions of test validity, given that these consequences are intimately connected to the test's interpretation, use, and effectiveness (Brennan, 2006). Some formulations of validity which are more squarely in the disciplinary mainstream accept that both intended and unintended consequences inform validity arguments for a test in the presences of systematically different outcomes for members of different population subgroups, *but only when the group differentials are rooted in flaws in the test design*, as the result of construct underrepresentation or construct-irrelevant variance (AERA/APA/NCME, 1999). Indeed, some researcher point to evidence that scores from a number of educational and psychological test have historically exhibited construct irrelevant bias for some groups of ethnic and racial minorities in the U.S. (Garcia & Pearson, 1994), and that in some cases, this kind of bias has contributed to a variety of negative outcomes, such as the disproportionate identification of student of color for special education services (Artiles, Harry, Reschly, & Chinn, 2002). Similarly, under current accountability requirements which mandate the reporting of scores for a number of student subgroups, there are concerns about the quality of inferences to be made from test scores of English Language learners (Abedi, 2004). Under some expanded conceptions of

test validity, these kinds of outcomes would provide evidence which might call into question the consequential validity of the interpretation and use of a given test.

However, a number of scholars have advocated for assessing the broader *social consequences* of the use of a given test when considering its overall validity (notably, Messick, 1989). Often, this view is advanced together with the assertion that educational systems are best thought of as complex dynamic systems, responding to assessment data over time in an adaptive fashion (Frederiksen & Collins, 1989). Fredricksen and Collins have suggested that the *systemic validity* of a given assessment or measure could be judged by the degree to which its use brings about desired systemic change and behavior (e.g. improved teaching and learning) or undesirable adaptive responses to the testing regime (e.g. teaching to the test which does not generate transferable knowledge). While considering these kinds of broad systemic responses as part of the test validation process is contentious within the research community, there is some evidence that this perspective has influenced policy documents, including those providing guidance for education policy at the federal level. Specifically, the U.S. Department of Education's *Standards and Assessments Peer Review Guidance* (2009), describes how states may meet the requirements of the No Child Left Behind Act of 2001 though adopting reliable and valid assessments. This document explicitly cited Messick (1989) in asserting that a state must consider the intended and unintended effects of the assessment in an ongoing validation process, including looking at both systemic outcomes (such as student grade-level retention) and mediating processes (including changes in teacher professional development. Research literature has documented a number of unintended mediating processes which have emerged as a response of school systems to test-based accountability (Cheng & Curtis, 2004; Jennings & Bearak, 2014; Herman, 2008). These organizational responses to assessments used for accountability purposes might be sorted into three larger categories of responses: (1) reallocation of instructional time – both between and within content areas – to reflect the content and rigor of the assessed curriculum; (2) teaching which focuses specifically on the features and formats of the accountability tests (i.e. "test prep"), and (3) cheating and gaming of the accountability system.

The influence of accountability assessment on instruction. A number of studies have documented that teachers adjust the form, content, and rigor of instruction to reflect accountability tests (Cheng & Curtis, 2004; Hamilton et al., 2007; Herman 2004, 2008; Jennings

& Bearak, 2014; Madaus et al., 1992; Stecher, Barron, Chun, & Ross, 2000). In some cases, this has been described as a desirable and intended response of policymakers designing a standards-based education system (Cheng & Curtis, 2004: Popham, 1987). For example, there is evidence of schools and teachers responding to changes in assessments in ways which reformers would interpret as desirable and envisioned by the theory of change of standards-based reform (e.g. Kortez, Barron, Mitchell, & Stecher, 1996; Stecher 2002, 2010). However, not all attempts of teachers to align their instruction with assessmen`t can be characterized as intended or desirable. Some critics of standards-based education reform and its implementation have pointed to assessments driving a "narrowing of the curriculum," citing cases of teachers allocating instructional time away from untested content areas to emphasize tested content areas (e.g. reallocating time from art to mathematics), or spending more time within a content area on those topics most likely to appear on state tests (Au, 2007). Additionally, teachers' adjusting of instruction to align with the form, content, and rigor of accountability testing is apt to result in diminished quality of learning when accountability tests measure knowledge and skills at very low levels of rigor and cognitive demand, as appears to be the case with the majority of accountability assessment utilized since the No Child Left Behind reforms (Darling-Hammond & Adamson, 2010; Yuan & Le, 2012).

Other organizational responses to accountability assessments: Test prep, accountability gaming, and cheating responses. While the ways in which assessment influences instruction may be more or less desirable, depending on the assessment and the context, there are a number of other documented responses to accountability testing which are in general less desirable and not accounted for in the theory of change of standards-based education reform. These include the use of instructional time for test prep activities and responses by teachers or schools to artificially raise proficiency rates through cheating and gaming the accountability system. In general, the kind of "teaching to the test" described as "test prep" is problematic in that teaching with an emphasis on item formats and other elements specific to a particular test is unlikely to produce knowledge and skills that generalize to other contexts, even when that context is another student assessment. As discussed above in the discussion of the phenomenon of test score inflation, research provides some evidence to suggest that classroom instruction focused on test prep might produce increases on scores used for accountability purposes, even though these

gains cannot be validated by scores from similar low-stakes assessments (Holcombe, Jennings, & Koretz, 2013; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz, 2005). Researchers have also documented a number of ways in which some schools have responded to particulars of the test-based accountability system, artificially boosting aggregate scores and proficiency rates. These kind of responses include a disproportionate focus on students close to proficiency cutoffs (Booher-Jennings, 2005), the manipulation of the testing pool (Cullen & Reback, 2006; Figlio & Getzer, 2006; Jacob, 2005), and cheating (Jacob & Levitt, 2003)

Value-added Measures

Value-added Modeling: The Average Residuals Approach

As discussed above, the student scores from most tests used for accountability purposes reflect a number of influential factors from outside the classroom. Individual scores and score growth vary tremendously within teacher, reflecting individual student differences in ability, out-of-school influences, and prior learning. As mentioned previously, the influence of the teacher is estimated to account for a relatively small percentage of the total variation in the growth of student achievement scores, about 10 percent (Goldhaber, Brewer, & Anderson, 1999). For this reason, sophisticated statistical models are necessary to isolate and estimate teacher contributions to student learning based on very general assumptions about the many factors which contribute to what students know and know how to do. In the *generalized cumulative effects model* of student learning (Boardman & Murnane, 1979; Guarino, Reckase, & Wooldridge, 2015), that which a student knows and is able to do is a function of all prior schooling experiences, all prior out-of-school experiences, and time-varying and time invariant individual characteristics. This general model can be expressed with the following equation:

$$L_{it} = f_t(E_{it}, \dots, E_{i0}, X_{it}, \dots, X_{i0}, c_i, u_{it}) \quad [1]$$

In this formalization, the learning of student *i* at time *t* is a function of all previous school-related inputs or experiences $(E_{it}, \dots, E_{i0})$ and all previous out-of-school inputs or experiences (including time-varying individual characteristics)$(X_{it}, \dots, X_{i0})$, individual time invariant characteristics (c_i) and unobserved exogenous shocks factors (u_{it}). One model which is

used to describe a reasonable and approximate relationship between these variables is a *general linear formulation* of the cumulative effects model (Guarino et al., 2015), which assumes that these variables contribute to learning in time *t* in a linear, additive fashion, with coefficients distributed to each lagged variable:

$$L_{it} = \beta_0 E_{it} + \beta_1 E_{it-1} + E_{i0} + \cdots + \gamma_0 X_{it} + \gamma_1 X_{it-1} + \ ... + \gamma_t X_{i0} + \ \eta_t c_i + \ u_{it} \quad [2]$$

The coefficients on the lagged variables (corresponding to times *t-1, t-2, …, 0*) can be interpreted as the degree to which the effects of these prior inputs or experiences fade-out over time. One common simplifying assumption frames the influence of all time-varying inputs, whether they be out of school or in school experiences, as following a geometric decay function, such that the influence of previous years' contributions to current learning shrinks by a fixed proportion λ each year (where $0 < \lambda < 1$). If this assumption is applied to the above equation, and we assume that school and outside-school factors decay at the same rate, then:

$$\beta_S = \lambda^S \beta_0, \gamma_S = \lambda^S \gamma_0 \quad [3]$$

where *S* is the number of years since the student was subject to the schooling or out-of-school learning experience. This assumption allows the term λL_{it-1} (less individual time-invariant contributions (*c*) and unobserved time-variant factors (*u*) for time *t-1*) to be substituted for the cumulative influence of all prior years' out-of-school and in-school experience on the current level of student learning, along with. This substitution results in the following equation:

$$L_{it} = \lambda L_{it-1} + \beta_0 E_{it} + \gamma_0 X_{it} + (\eta_t c_i - \lambda \eta_{t-1} c_i) + \ (u_{it} - \lambda u_{it-1}) \quad [4]$$

Or,

$$L_{it} = \lambda L_{it-1} + \beta_0 E_{it} + \gamma_0 X_{it} + \pi_t c_i + e_{it} \quad [5]$$

if we define e_{it} as equal to $(u_{it} - \lambda u_{it-1})$. Because we are unable to measure student learning at time *t* without error, teacher contributions to student learning are usually estimated using achievement scores at time *t* (A_{it}).

One of the most common approaches to estimating these teacher contributions – the average residual approach (Guarino et. al, 2015, Kane & Staiger, 2008) – regresses student test scores on prior year tests scores (i.e. regresses A_{it} on A_{it-1}) and other available proxies for out of school influences (such as student socioeconomic status) to obtain a student level residual for each data point, $\hat{v}_{it}$. In this approach for student *i* assigned to teacher *j* in year *t*, the first step estimation model would be:

$$A_{it} = \lambda A_{it-1} + \gamma_0 X_{it} + v_{it} \quad [6]$$

where X_{it} is a vector for proxies of out-of-school influences. Here, the residual consists of three components:

$$v_{it} = \left(\gamma' E_{jt} + u_{it} - \lambda u_{it-1}\right) \quad [7]$$

Where E_{jt} is the effect on student learning for student *i* of teacher *j*, and *u* represents the exogenous shocks to learning at time *t* and *t-1*. In the second step of the estimation, these residuals ($\hat{v}_{it}$) are then averaged within teacher to obtain an estimate of average teacher contribution to student learning ($\hat{E}_{jt}$).

Assumptions. Reardon and Raudenbush (2009) outline a series of assumptions implicit in value-added modeling, in general. These authors specify two assumptions as "defining assumptions," necessary to interpret the estimated statistics as an average or expected causal effect for assigning any student in the population of interest to a given teacher. In order to interpret the derived statistics in this way, we need to assume that (1) it is conceivable or theoretically possible that any student in the population of interest could possibly be assigned to any teacher in the population, and (2) that the effect of teacher assignment on a given student is independent of other students' assignments to teachers. While the first assumption may be challenged by the realities of student assignment to schools and teachers (given neighborhood zoning policies and between and within school sorting on prior achievement), it is important that alternate scenarios of students to teachers be at least theoretically conceivable in order for the inference of causal effect to be theoretically meaningful.

The second assumption – that student outcomes are independent of other students' assignments has been described as the *stable unit treatment value assumption* (SUTVA; Reardon

& Raudenbusch, 2009; Rubin, 1986), and can more simply be thought of in the context of education research as the assumption that classroom-level peer effects are negligible. This assumption is important in defining the value-added estimate as the causal effect of assignment to a given teacher, as opposed to student assignment to a given set of teacher and students. In theory, classroom peer composition may contribute to student learning directly (e.g. through activities such as peer tutoring) or indirectly (e.g. through other students causing distractions or disruptions, asking particularly helpful or interesting questions, or otherwise influencing teachers' choice of instructional practices and curricula) (Henry, Rose, & Lauren, 2014; Reardon & Raudenbusch, 2009). If the presence or absence of other students contributes in a non-negligible way to average student performance on the achievement measure employed, then the coefficients estimated by the statistical models will not disentangle but instead conflate the effect of teacher assignment and peer group.

While the SUTVA assumption concerns interpretation of estimated parameters and teacher-level contributions to student learning, Reardon and Raudenbush (2009) delineated another set of assumptions required to allow for unbiased estimation of the parameters of interest from the available observed data. These include assumptions that:

- Test scores are measured on an interval scale;
- Causal effects do not vary as a function of student background;
- Either there is enough overlap and diversity in the distribution of students to each kind of teacher to support the extrapolation of average treatment effects to all kinds of students, *or* the functional form of the model correctly specifies the likely student outcomes for the kinds of students who are not assigned to a given teacher, and;
- Any confounding relationship of assignment to teacher and end-of-year test score outcome is taken into account through control variables included in the model;

The last assumption hinges on the specifics of the estimation model and the control variables utilized. Guarino, Reckase, and Wooldridge (2014) addressed some of the assumptions which apply to the average residual approach. A number of these assumptions can be seen through looking more closely at the models in both the first- and second steps of the estimation (Equations 6 and 7), repeated here:

$$A_{it} = \lambda A_{it-1} + \gamma_0 X_{it} + v_{it} \quad [8]$$

$$v_{it} = \left(\gamma' E_{jt} + u_{it} - \lambda u_{it-1}\right) \quad [9]$$

Here, we see an implicit assumption that, conditional on student time invariant characteristics or demographics (X_{it}), prior achievement (A_{it-1})and assignment to teacher (E_{ijt}) are independent. Guarino and colleagues (2014) described this assumption or constraint as one of the shortcomings of the average residual effect, that it does not control adequately partial out the relationship between teacher assignment and lagged test scores and other control variables, resulting in biased and inconsistent estimates when prior achievement varies systematically by teacher. For this reason, Guarino and colleagues have voiced their preference what they term a *dynamic ordinary least squares* (DOLS) approach, which estimates teacher value-added in one step, through including teacher assignment dummy variables in the first step equation. However, as this average residual approach is frequently estimated in research – notably the estimates for value-added generated by the MET Project researchers (c) – additional teacher- or class-level controls are added to the first step equation in order to account for some of this sorting or potential confounding correlation between student prior achievement and assignment to teacher, as in the following equation:

$$A_{it} = \lambda A_{it-1} + \theta \bar{A}_{ijt-1} + \gamma_0 X_{it} + \alpha \bar{X}_{ijt} + v_{it} \quad [10]$$

where $\bar{A}_{ijt-1}$ represents and average prior student achievement of students and $\bar{X}_{ijt}$ represents mean student demographic characteristics, both averaged within teacher j. This might be interpreted as adequately controlling for assignment of students to teachers based on a students' prior achievement.

Equation 7 also highlights the fact that, in order for the estimate of teacher causal effects to be estimated without bias, assignment to teacher should be independent of exogenous shocks to learning at in both the current and prior year (i.e. u_{it} and u_{it-1}) respectively. Concrete example senarios cited in the literature could be family decisions to compensate for student assignment to a particularly poor teacher with after school private tutors, although it might be difficult to conceive of these compensatory behaviors being systematics enough to introduce substantial bias into the estimates.

Alternate Estimation Methods

For the purposes of this dissertation, approaches to modeling and estimating teacher contributions to student learning can be sorted into three larger categories: (1) univariate response value-added models, (2) multivariate response value-added models, and (3) student growth percentiles. The average residual approach described above is a univariate response value-added model: common variations within this class of models include those which estimate teacher effects with a single step (DOLS or teacher fixed-effects models), those which use the gain score as a dependent variable (assuming complete persistence/no decay of prior learning), the application of Bayesian "shrinkage" to estimated teacher effects, and hierarchical linear approaches to explicitly model nested data (Guarino, Reckase, & Wooldridge, 2014; McCaffrey, Lockwood, Koretz, & Hamilton, 2004). In addition to their utilization in research, measures from the univariate response family of teacher-value added models have been used to inform policy in a number of state-level departments of education (including Minnesota, North Dakota, South Dakota) and school districts (Atlanta, Chicago, Los Angeles; Hillsborough County, FL; Milwaukee, Tulsa, and New York City, Madison, WI;, and Washington, DC) (Value-added Research Center, 2015; Walsh & Isenberg, 2015)

The second family of value-added models, consisting of multivariate longitudinal models, estimates covariance matrices from the joint distribution of student test scores from multiple years and across multiple subject (McCaffrey et al., 2004). These covariance matrices are then used to estimate individual growth trajectories for each student, modeling teacher effects as a measureable and persistent deflection in the individual estimated growth trajectory. Although these models do not typically control for student demographic characteristics, they do estimate relationships across many subject area tests and across multiple years, requiring larger and richer data sets than the univariate value-added models described above. These multivariate longitudinal models – also referred to as *layered* models – are utilized by state-level agencies in Tennessee, Ohio, Pennsylvania, North Carolina, and South Carolina, as well as in a number of districts in other states (SAS Institute, 2015).

The third class of models described here – student growth percentile (SGP) models – are used to generate teacher-level statistics which describe student growth. While SGP scores are in practice treated and interpreted as being similar to teacher-value added estimates, their

interpretation is theoretically district from value-added estimates (Goldhaber, Walch, & Gabele, 2014). In contrast to value-added models, statistics generated from SGP models are not designed to produce causal estimates, but are instead designed to produce statistics which are descriptive and more easily communicated to and interpreted by educators and other stakeholders (Betebenner, 2007; Goldhaber et al., 2014). In effect, growth percentile models calculate the achievement percentile for a student in a given year relative to all other students with an identical scores in the prior year, generating a *student growth percentile* (SGP). Students are then matched to teachers, and the within-teacher mean or median student growth percentile (MGP) is used as a statistic to describe the average within-teacher growth on state assessments. Student growth percentile measures have been or are slated to be employed by at least 16 educational agencies at the state level, including Arizona, Arkansas, Colorado, Georgia, Hawaii, Indiana, Kentucky, Massachusetts, Mississippi, New Hampshire, New Jersey, Rhode Island, Virginia, Washington, West Virginia, and Wyoming (Walsh & Isenberg, 2015).

Issues of Reliability and Validity for Teacher Value-added Estimates

Reliability. A number of studies have examined the stability of teacher value-estimates over time (e.g. Arronson, Barrow, & Sanders, 2007; Ballou, 2005; Goldhaber & Hansen, 2013; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009). These analyses typically utilized a number of approaches to describing the variation in teachers' value-added estimates over time. One approach focuses on value-added as a continuous measure, using statistics such as correlation coefficients or describing the proportion of variance shared by repeated estimates of teacher value-added over time. A number of these analyses also described year-to-year consistency in teacher value-added measures through reporting a different metric: change in quintile membership over time. Use of this metric for reliability is motivated by the justification that it is potentially more policy relevant, given actual and purposed potential sanctions and rewards may be meted out to teachers based on their classification to these categories.

Studies that report correlations – most frequently Spearman rank correlations – report a wide range of year-on-year correlations of estimated teacher value added. McCaffrey, Sass, Lockwood, & Mihaly (2009) found that year-to-year correlations range from 0.2 to 0.5 for elementary teachers and 0.3 to 0.7 for middle school teachers. More recently, Goldhaber and

Hansen (2013) fount correlations on the higher end of these ranges with a correlation of 0.55 for adjacent year estimates, and an upper-end of 0.65 for estimates corrected for error variance using empirical Bays adjustment. By these measures, teacher value-added reliability is comparable to and perhaps superior to objective measures of job performance in other professions: a meta analysis of 22 studies found the average year-to-year correlation of objective occupational performance measures to be 0.37 (Sturman, Cheramie, & Cashen, 2005).

A number of opinion pieces and literature reviews which discuss the appropriateness of using teacher value-added estimates for both research and policy purposes typically cite a lack of high year-over-year correlation as a point of concern (American Statistical Association, 2014; Baker et al., 2010; Ballou & Springer, 2015; Corcoran & Goldhaber, 2013; Glazerman et al., 2010; left Guarino, Reckase,& Wooldridge, 2015). While disagreeing on failing to specify what would constitute an acceptable level or reliability or stability for accountability purposes, most of these commentaries agree on the usefulness of these measures for research to describe and provide insight into systemic educational processes In general, most of these authors agree that if used to inform staffing decisions, value-added estimates should ideally be combined over multiple years and be utilized as part of an overall comprehensive evaluation system which includes other complementary performance measures.

Validity of value-added estimates. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014) identified value-added measures as an "accountability index," derived from multiple data sources and involving complex statistical modeling (p. 206). The *Standards* asserted that accountability indices require additional evidence for validity for their use and interpretation, beyond the evidence for validity of the individual measures from which they are derived. That being said, it is still important to keep in mind the validity and limitations of the assessments of student achievement as measures of student learning. Teacher-value added models seek to model student achievement data in such a way as to isolate and provide an estimate of each teacher's average contribution to student learning on achievement tests. However, as single measure of teacher effectiveness derived from students' achievement tests, it is important to acknowledge that these estimates are limited to teacher contributions to learning adequately measured by those tests. As such, these estimates are will not may not be able to measure all beneficial teacher effects on students, such as the indirect effects of a veteran

teacher mentoring other teachers (Kupermintz, 2003) or teacher's more direct effects on unmeasured social and behavioral skills (Jackson, 2013; Jennings & DiPrete, 2010).

Value-added and teacher characteristics. Given that caveat, a search for evidence of validity of value-added measures as indicators of teacher effectiveness could begin with a search for confirmatory relationships with other variables: Do value-added estimates of teacher effectiveness tend to correlate positively with measures from other constructs we would expect to show these relationships, a priori? While salary schedules have typically rewarded teachers with more experience and advanced certification or degrees, these indicators typically are not always positively correlated with teacher value-added scores. Specifically, a number of studies have shown that teacher value-added estimates tend to increase only during the first three- to five years of experience, plateauing after that period (Cavalluzzo, 2004; Hanushek, Kain, O'Brien, & Rivkin, 2005; Kane, Rockoff, & Staiger, 2008; Rockoff, 2004). On average, teachers with master's degrees have not been found to have higher value-added estimates (Clotfelter, Ladd, & Vigdor, 2006; Rowan, Correnti, & Miller, 2002). However, in the field of mathematics teaching, there is some evidence to suggest that teachers tend to have higher value-added scores when their degree major or minor was in mathematics (Aaronson, Barrow, & Sander, 2007; Monk, 1994; Goldhaber & Brewer, 2000) or when observations of their teaching provides evidence of greater pedagogical content knowledge (Hill, Kapitula, & Umland, 2011). However, estimates of teacher pedagogical content knowledge as measured by written assessments tend to exhibit only very weak relationships with value-added scores (Gitomer, Phelps, Weren, Howell, & Croft, 2014; Hill et al., 2011). Additionally, teachers' scores of general knowledge and verbal ability have been associated with gains in student learning (Ferguson & Ladd, 1996; Rice, 2003).

Value-added and measures of instructional quality. Compared to the measures of the teacher characteristics described above, we might expect that more direct measures of instruction might have stronger relationships with teacher value-added estimates, given that they are more proximal to student learning. Variables in this category might include measures of teachers' coverage of the curriculum, measures of quality of instruction on rubrics for observing classroom teaching, and subjective ratings of teachers by principals (which are, presumably, informed by both teacher characteristics and classroom observations). Polikoff and Porter (2014) analyzed teacher surveys results to derive quantitative measures of the degree to which they cover their

state standards in mathematics. These measures of teachers' instructional alignment to standards demonstrated very small but statistically significant correlations with teacher value-added estimates (0.16, $p<0.05$). In contrast, Le and colleagues (2009) did not find their measures of teachers' curricular coverage to be a significant predictor of student gains.

Scores from researcher generated scores on a number of classroom observational rubrics showed relatively weak positive correlation with one year value-added estimates, ranging from 0.03 to 0.18 (Polikoff, 2014). Other researchers (Kane & Staiger, 2012) utilized the same data and measures but attempted to isolate an "underlying value-added score," using two years' of teacher value and factoring out year-to-year fluctuations in teachers' estimated value added, arriving at a more stable or persistent component of teacher value-added (pp. 39-40) . While these "underlying" value-added showed somewhat larger correlations with observational scores (ranging from 0.09 to 0.34), they were still not statistically significant at conventional levels (here, p-values not less than 0.10) (Kane & Staiger, 2012). In general, these measures were more effective at distinguishing teachers at the extreme ends of the value-added distribution than distinguishing differential teacher value-added between teachers within the middle of the distribution (Kane & Staiger, 2012).

Finally, there is some evidence that, on average, principals can draw upon their knowledge of teachers and the school context to predict which teachers will contribute more or less to students' gains on assessments. Compared to the correlations from classroom observations and teacher value-added in the MET data, Jacob and Lefgren (2008) found higher correlations between principals' subjective evaluations and teachers' value-added estimates, ranging from 0.18 to 0.55, and with greater predictive power ($p<0.05$). As in the analysis of observation measures (Kane & Staiger, 2012), these principal estimates of teacher effectiveness were much better and distinguishing teachers at the extreme ends of the distribution than those in the middle of the value-added distribution.

Value-added and measures of instructional quality across different student assessments. A small number of studies look at the way in which changing the student assessment alters the relationship between estimated teacher value-added and measures of instructional quality (Grossman, Cohen, Ronfeldt, Brown, 2014; Le et al 2009; Polikoff, 2014; Walkington & Marder, 2014). Most of these analyses have utilized data from the Measures of

Effective Teaching (MET) Project (Kane & Staiger, 2012), which is well suited to this kind of analysis, given that this project's dataset allows for the comparison of scores from multiple instruments for measuring instructional quality through classroom observations and includes value-added measures from both state accountability tests and commercially available tests with open response items. These kinds of analyses are of particular interest for this dissertation, given that my analysis is also looking for changes in the relationship between measures of ambitious instruction in mathematics and teacher value-added estimates after states transition to assessment aligned to college- and career-ready standards.

Some of these analyses suggest that assessments vary in the degree to which they are sensitive to differences in teachers' instruction. Using a sample from six school districts in six different states, Polikoff (2014) analyzed the relationships between teacher value-added estimates and several classroom observation instruments and found that, for a given measure of instructional quality, correlations with value-added varied substantially across these districts. Because these districts used different student assessments, Polikoff concluded that these assessments vary in the degrees to which they are sensitive to instruction. Also utilizing MET Project data, Grossman and colleagues (2014) found that value-added scores from more rigorous, open-response tests are more highly correlated with scores on the PLATO classroom observations instrument than were value-added scores from state assessments. Furthermore, these researchers found that this stronger relationship is driven by the subscale of the observation instrument which measured students' participation in more intellectually challenging activities and discussion. Grossman and colleagues concluded that, in general, assessments likely vary in the degree to which they are sensitive to and reward more ambitious and cognitively demanding forms of teaching and learning.

There is some limited evidence that assessments in mathematics also vary in the degree to which they are sensitive to more ambitious instruction (Le et al.2009, Walkington & Marder, 2014). Walkington and Marder (2014) analyzed the relationship between scores from the UTeach Observation Protocol (UTOP) for mathematics and value-added on both the state assessment and the Balanced Assessment in Mathematics for 249 teachers in grades 4 through 8. These authors concluded that, overall, the Balanced Assessment in Mathematics rewarded inquiry-style instruction more than the state assessments. However, they also found that even

holding the assessment constant, the relationship between value-added and the inquiry-oriented instruction varied substantially by grade, even for the Balanced Assessment in Mathematics. Le and colleagues (2009) found that while a number of ambitious math teaching practices were negatively correlated to student gains on multiple choice tests measuring students' knowledge of and facility with mathematical procedures, these same practices were positively correlated with student gains on multiple choice tests measuring problem-solving, with even stronger positive correlations for tests with open-ended response items.

Convergent validity and student demographic characteristics. Finally, some research has also probed the relationships between teacher value-added estimates and student-level variables. Some research has found that teacher value-added estimates tend to correlate positively with class-average prior achievement scores and negatively with class-level background variables which might identify an at-risk population (e.g. proportion of English language learners or proportion of low-income students) (i.e. Hill et al., 2011), which led some researchers to question if teacher-value estimates are more of a reflection of student-demographics and school-level effects. However, much research also suggests that teachers with lower indicators of teacher quality are disproportionately found in schools with higher concentrations of at risk students (e.g. Boyd, Lankford, Loeb & Wyckoff, 2003; Clotfelter, Ladd, & Vigdor, 2004). Given the evidence of the systematic and unequal distribution of teacher quality by student demographic, the negative correlations between teacher value-added scores and some student characteristics might be interpreted as evidence for the convergent validity of value-added measures. Finally, a handful of published studies have found that some evidence that teacher-value added measures demonstrate a degree of predictive validity. These analyses have found that students who have been taught by higher value-added teachers are more likely to experience more positive outcomes years later, including reduced teen pregnancy and increased college going and earnings (Chamberlain, 2013; Chetty, Friedman, & Rockoff, 2014).

Consequential validity of value-added measures. Theory and research in incentives and personnel evaluation have established that, in order for personnel evaluation systems to work effectively, evaluation criteria should be transparent (van Herpen, van Praag, & Cools, 2003). This is to say that the most effective employee feedback and evaluation systems employ criteria or measures which are unambiguous and easily understood. As described above, students'

scores on most commonly used assessments are sensitive to variation from a number of different sources, such that sophisticated statistical techniques used to isolate the relatively small proportion of variation in these scores which is attributable to teachers' contributions to learning. Unfortunately, neither the process nor the product of value-added estimation is transparent to the majority of practitioners, making problematic their value and implementation for accountability purposes (Ballou, 2002).

Working from theory, there are a number of ways in which the use of teacher-value added estimates for evaluating, rewarding, or sanctioning teachers could potentially be problematic (Baker et al., 2010), resulting in negative unintended consequences which influences the perceived consequential or systemic validity of their uses this way. Value-added measures are derived from student assessment scores. Accordingly, some of the potential unintended consequences which are associated with high-stakes testing in general are also associated with teacher evaluation which incorporates test-derived value-added estimates. However additional potential negative unintended consequences of implementations of value-added to policy stem from some of the real or perceived ways in which subgroups of students perform relative to each other, and also from the way in which the process of estimation essentially ranks teachers against each other, as opposed to judging their performance relative to fixed objective criteria. Baker and colleagues (2010) suggested that incorporating value-added estimates into teacher evaluation might provide disincentives for teachers to work with groups of students whom they perceive as unlikely to demonstrate growth, including individual students experiencing a idiosyncratically difficult year, enduring experiences such as a recent move, illness, or parents' divorce. Furthermore, because value-added methodologies describe teacher effectiveness relative to other teachers, there may be disincentives to collaborate. These disincentives may be particularly problematic in light of empirical research evidence which points to the importance of between-teacher trust, collaboration, and peer learning in effective schools (Bryk & Schneider, 2002; Goddard, Goddard, & Tschannen-Moran, 2007; Jackson & Bruegmann, 2009).

Some empirical work affirms that these negative unintended consequences have been observed in some cases. Collins (2014) reported a number of negative unintended consequences in a U.S. school-district where value-added estimates had been used to reward teachers, including: teachers avoiding or feeling penalized for teaching students they perceive as less

likely to show growth (in this case, gifted students, special education students, and English-learners in their first year of transition to English-only education); perceptions of principals rewarding or punishing teachers based on value-added scores through their assignment of students to teachers; teachers seeking to influence class makeup through negotiations and favor-seeking with administrators; cheating, gaming, or teaching to the test; lack of incentive to collaborate, and; lowered teacher morale, including for teachers who teach untested grades and subject areas and are not eligible to receive bonuses. It is important to note that these negative effects will not inevitably occur teacher-value added scores are or may be used for evaluation and compensation decisions, and that much depends on the particulars of context and implementation. For example, there is some suggestion that the use of teacher value-added estimates might be structured into group-level evaluations and incentives in order to encourage rather than discourage teacher collaboration, although recent experiments with these kinds of incentives have not found significant positive effects (Marsh et al., 2011; Springer et al., 2012)

Variation in Teacher Value-added Introduced by Changing the Estimation Approach

In practice, different estimation methods should yield different estimates of teacher effectiveness using identical data sets. A number of analyses have utilized real or simulated datasets to quantify and describe the difference in teacher value-added estimates between different methods. These analyses typically report the degree of difference between these results from different estimation approaches by reporting a correlation statistics. These analyses also typically simulating data to measure correlations between estimated teacher value-added and the "true" teacher effect specified by the simulated data set and/or report on the different correlation statistics between the results of estimating teacher value added from real data. Analyses using simulated data which allow them to compare the ability of different modeling approaches to estimate the "true teacher effect," even as some of the assumptions about the sorting of teachers and students is violated (Henry, Rose, & Lauren, 2014; Guarino, Reckase, & Wooldridge, 2014). Henry, Rose, and Lauren (2014) employed simulated data to find that estimated teacher effects correlate highly or very highly with the true effect (0.87 to 0.93) assuming random assignment of teachers and students, with most violations of random assignment showing still high correlations

(0.72 to 0.91).[6, 7] Guarino and colleagues (2015) simulated data in elementary school scenarios (i.e. fewer students per teacher) and find that moderate to high correlations under scenarios of random assignment, as well as under most violations of random assignment of teachers and students (ranging from 0.52 to 0.91, with an average of 0.81).[8]

Other studies have used large real data sets to estimate teacher effects using different models, although in these cases, the "true" teacher effect is unknown, and values from different estimation approaches are compared with each other (Goldhaber, Walch, & Gabele ,2014; Sass, Semykina, & Harris, 2014). Using statewide data, Goldhaber, Walch, and Gabele (2014) compared single stage, teacher-fixed effects estimations with percentile growth measures and find correlations ranging from 0.92 to 0.93 for math scores. However, these authors also find that the inclusion of school level-fixed effects (effectively comparing teachers within schools) produces estimates which depart to a greater degree from the other value-added and percentile growth measures without school fixed effects (correlations ranging from 0.61 to 0.65). Sass, Semykina, and Harris (2014) also used administrative data and find that, in general, the rank correlations of teacher effects using five ordinary least squares specifications were in general high or very high (0.69 to 0.94).[9]

However, these analyses are also concerned with the policy ramifications of using one model over another, and the extent that they could result in *miscategorization* of teachers. In practice, teacher value-added estimates have been used to reward or sanction teachers who fall

[6] In characterizing these correlations qualitatively, I have tried to adhere to guidelines suggested by Hinkle, Wiersma, & Jurs (2003) for using terms such as *very high correlation* (0.90 +), *high correlation* (0.70-0.90), *moderate correlation* (0.50-0.70), *low correlation* (0.30 to 0.50), and *little if any correlation* (<0.30).

[7] Correlations reported here are for simulations of middle-school teachers teaching multiple classrooms of students. Compared to the middle school scenario, the correlations Rose and Henry report for elementary school scenarios/class sizes are slightly higher under assumptions of random assignment (0.91 to 0.96) and lower under violations of random assignment (0.65 to 0.86)

[8] Results from Guarino, Reckase, and Wooldridge (2014) reported here exclude correlations from the student fixed effects on gain score and an instrumental variables/Arellano and Bond approach, both of which (which performed substantially worse in these simulations (correlations ranging from 0.47 to 0.74, average correlation = 0.58). Results also exclude one of the 9 combinations of violations to random sorting assumptions; a scenario with student groupings based on prior-year scores with more effective teachers assigned to students more likely to resulted in a wide variation of correlations of estimated teacher effect and true teacher effect (-0.44 to 0.87).

[9] Sass and colleagues (2014) also analyzed the correlations between and among two less widely used estimators: first-differencing estimators and (as in Guarino, Reckase, and Wooldridge, 2014) students-fixed effects estimators. The correlations of estimates among the first-difference models were high or very high (0.84 to 0.99) and only moderate between the two student fixed effects models (0.65). However, results from within these three classes of models (ordinary least squares, first difference, and student fixed effects) were typically much higher than the correlations between results from different types of estimators (0.19 to 0.58).

into the highest or lowest quintile, respectively. As a result, these research analyses often quantify not only the correlation of the point estimates from different models, as described above, but also the rates of teachers who are classified to fall in different quintiles as the estimation model changes.[10] However, this dissertation analysis is more concerned with the point estimates from value-added models and their relationship, as a group, to features of classroom instruction, and the degree to which the relationship between certain features of classroom instruction and teacher-level contributions to student learning growth with the adoption of new and ostensibly more rigorous state tests. This analysis will necessitate estimating teacher-level contributions to student learning using a number of models to provide a check for the robustness of findings, given that there is some meaningful variation between the sets of estimates produced by different models.

Variation in Teacher Value-added Introduced by Changing the Student Assessment

At the same time, there have been a number of studies which have looked both the differences in teacher value-added ranks when (1) using different measures of student achievement and (2) using different models for estimating teacher value added (Lockwood and colleagues, 2007; Papay, 2011). These studies have found that, generally, teacher value-added estimates are more sensitive to changes in the student outcome measure than they are to changes of model specification.

Lockwood and colleagues (2007) utilized four years of longitudinal test score data from 3,387 during their studies in grade 5 through 8. They estimate teacher value-added scores using four univariate and multivariate approaches and five combinations of controls for each approach. They fit these models using two different measures of student achievement: the subtest of the Stanford 9 mathematics assessment used to measure students' proficiency with procedures and calculation, and the subtest of the same assessment which was designed to measure more complex problem-solving skills. Lockwood and colleagues reported that Pearson correlation coefficients between scores of different value-added specifications range from 0.49 to 1.00; while correlations which utilized the same model specification but different subtests as outcomes

were much lower, ranging from 0.01 to 0.46. These authors conclude that in this case, changing the assessment used to measure student achievement in substantially larger differences in value-added estimate than did changing estimation model specification.

Papay (2011) replicated this analysis with data from both subsets of the Stanford 9 assessment for math linked to up to 32,000 student-year records nested in 762 unique teachers. Papay concluded that his results generally support Lockwood and colleagues' (2007) earlier conclusions, affirming that: "Much more variation in teacher value-added estimates arises from the choice of outcome than the model specification" (p. 165). Furthermore, Papay interpreted his results as suggesting that these two tests do not appear to be measuring a unidimensional construct of mathematical knowledge, but that they instead measure two distinct dimensions of math knowledge.

This Analysis

This analysis will investigate the relationship between a classroom measures of ambitious teaching in mathematics and covariate adjusted student learning gains (i.e. teacher value-added) to determine the degree to which these measures of ambitious teaching are correlated to the estimates of teachers' average contributions to student learning as measured by these tests. The magnitude of the relationship, if any, between the measures of teaching instruction and the test-derived measures of teacher contribution to student learning will be interpreted in this analysis as the sensitivity of the assessment to instruction aligned with ambitious goals for student learning. This analysis will look at two districts and the NCLB-era assessments they used from study years 1 to 4 (i.e. 2007-8 to 2010-11).

In study years 5-7 our districts adopted career and college ready standards and end of the year accountability assessments aligned to these new standards. This analysis will also estimate the degree to which the relationships between the measures of the quality of teaching instruction and the test-derived measures of teacher contribution to student learning changed with the adoption of these new tests.

Research Relevance

The review of literature above provides evidence that, while assessments used for accountability are generally not good at measuring differences in the quality of instruction in general, and instruction which promotes ambitious goals for learning mathematics in particular, the extent to which tests measure the enactment of ambitious teaching varies substantially. The results of this analysis may provide additional insight into the degree to which these newly adopted tests, aligned to career and college ready standards for learning, can be used to measure ambitious teaching in mathematics.

Policy Relevance

The review of literature also describes findings about how tests influence instruction and the importance of assessments aligned to the educational system's explicit goals for teaching and learning identified by the system in order to realize the logic and benefits of standards-based education. Results from this dissertation analysis may inform policy, in that they may give an indication as to the extent to which these tests (1) are aligned with ambitious goals for teaching and learning math, and (2) can be used to measure and reward ambitious teaching and learning. In an analysis with similar research questions, Polikoff (2014) found the relationship between quality of instruction (as measured by classroom observation) and teachers' estimated value-added to be so weak as to cast doubt on the likelihood that value-added estimates can be used to inform instruction and instructional decisions in any meaningful way. It may be that, if these student assessments cannot be used to derive valid measures of ambitious teaching, then their use needs to be limited to purposes for which these test results are valid measures. Furthermore, given the logic of standards-based education reform and the influence of assessments on instruction more generally, policy makers who hope to promote ambitious goals for teaching and learning mathematics may need to advocate more strongly for student assessments which effectively measure and promote this kind of teaching and learning.

CHAPTER III

DATA

Study Setting

Data for this analysis comes from the Middle-school Mathematics and the Institutional Setting of Teaching (MIST) Project, which followed four large urban districts and their policies and teacher supports for implementing inquiry-based curriculum and instruction in middle school mathematics. These districts were in some ways atypical, in that they responded to federal-, state-, and local level accountability pressures to raise student test scores in middle school math by moving to a rigorous inquiry- and discussion-based approach to mathematics at these grades, supporting this policy with substantial investment in teacher development (Cobb & Smith, 2008). An examination of some key descriptive variables from our districts illuminates the context and scope of some this challenge (Table 1). Like many large urban districts, our partnering districts enroll high percentages of low-income and minority students. Specifically, three of the four districts enroll a proportion of English language learners that is much higher than the national average for large urban school districts. Furthermore, each of our four partnering districts consists of over 100 schools. In these expansive organizational contexts, improving the quality instruction in classrooms across the district – or indeed, any significant organizational change – becomes an issue of turning a metaphorical "battleship" (Olszyk & Kessler, 2008; Weinstein, 1993).

Table 1. Comparison of MIST participating school districts with national universe of all school districts and all urban school districts. Figures based on PreK-12 enrollments. Source: Common Core of Data, 2007. Large urban school district defined here as one located in a "City, Large Territory inside an urbanized area and inside a principal city with population of 250,000 or more."

	Dist. A	Dist. B	Dist. C	Dist. D	US School District (avg.)	All US Large Urban School Districts (avg.)
Schools	100	150	250	200	7.0	60.6
Teachers	2,000	5,000	12,000	6,000	220	2,201
Students	30,000	80,000	160,000	100,000	3,469	36,220
% Special Ed.	20	10	10	15	13.7	10.1
% ELL	20	30	15	5	17.1	11.8
% African American	40	25	30	40	7.8	24.1
% Hispanic	15	60	65	5	11.3	39.6
% FRPL	65	70	85	55	38.3	61.0

Note: statistics for partnering districts are rounded in order to maintain district anonymity

Within each of the four districts, six to ten middle schools were selected purposefully to construct a sample of middle schools which reflected the school-level variation of student demographics and achievement within each district. At each school site, three to five mathematics teachers were chosen randomly from a school roster by our researchers and recruited for participation. When a teacher left the school or study, another teacher from the same school was chosen at random and recruited to maintain the same number of participants. This study will focus on a longitudinal sample of two of these district – Districts B and D – which continued participating in data collection for a total of seven years. This provides the opportunity for the study of two relatively distinct periods of standards-based accountability. In Years 1-4, both districts worked under assessments aligned to No-Child-Left-Behind era standards. In Years 5-7, both districts were required to adapt to a shift in content standards and assessment aligned to college and career ready goals, as dictated by state-level policy.

The Teachers

Across the two focal districts for this study, middle-school mathematics teachers participating in our study have some commonalities (Table 2, Table 3). In both districts, over 65 percent of teachers are women; on average, these teachers have between eight and nine years of teaching experience, although the distribution is right skewed in both cases, such that the median years of teaching experience in both districts is closer to five years.

However, there are also some notable differences between participating teacher from these two districts, especially in regards to educational attainment and racial or ethnic background. Teachers from District D were more than twice as likely to possess a master's degree than teachers in District B, which is most likely attributable to differences in teaching certification requirements between the two states, with District B's state requiring that teachers obtain master's degrees within ten years of beginning their teaching careers (Morgen, 2017). Participating teachers in District B were more likely to be African-American, Hispanic, or American Indian, with teachers in District B more likely to be white. However, in both districts, the majority of teachers were white, though this majority was much slimmer in District B (where 51 percent of participating teachers were white) than in District D (where 80 percent of participating teachers were white).

Table 2. Descriptive statistics of demographic characteristics of teachers in the analytical sample, compared to district student population.

	District B	District D	District B	District D
	Teachers		Students	
Female	65.8%	67.9%	50%	50%
African-American	31.2%	13.6%	25%	40%
Asian-American	0.9%	0.5%	<5%	<5%
White	51.1%	84.8%	10%	50%
Hispanic	16.3%	0.5%	60%	5%
American Indian	4.5%	0.5%	<5%	<5%
Pacific Islander	0.0%	0.0%	<5%	<5%

Table 3. Descriptive statistics of certification, educational attainment, and years of experience of teachers in the analytical sample, compared to district student population. (Note: numbers reported here are from final analytical sample, as described in the results section)

	District B	District D
Certification		
Full Certification	92.8%	87.5%
Partial Certification	5.4%	12.0%
No Certification	1.8%	0.5%
Educational Attainment		
Associates Degree	1.0%	0.6%
Bachelor's Degree	66.0%	28.7%
Master's Degree	32.5%	69.1%
Doctoral Degree	0.5%	0.6%
(Degree unknown)	0.0%	1.1%
Years of Teaching Experience		
1 Year	8.6%	13.9%
2 to 5 Years	41.0%	39.0%
6 to 10 Years	21.2%	18.7%
11 to 15 Years	8.1%	6.4%
16 to 20 Years	9.5%	5.9%
More than 20 years	8.1%	13.9%
Average Years Exp.	8.4	8.8
Average rate of attrition from school[11]	22.1%	31.9%
Average rate of attrition from the district[12]	22.4%	16.0%
Unique Teachers	119	125
Teacher-Year Observations	226	211

Teacher Observational Measure

A team of researchers at the University of Pittsburgh developed the Instructional Quality Assessment (IQA, Boston, 2012; Boston & Wolf, 2006; Matsumura et al, 2006), building upon earlier work articulating frameworks for assessing the rigor of mathematical tasks and the

[11] These attrition rates are for years 1 through 4 of the study only. There is a large discrepancy between the attrition rate from the school and attrition in these years in District D due to teachers being in a *restart* or *turnaround school* where school leadership and staff in a chronically failing schools are disbanded and typically rehired in similar capacities in other schools in the district. See Duke (2012) for details on this kind of school turnaround under No Child Left Behind guidance and statutes, or see Rosenquist, Henrick & Smith (2013) for more on teacher attrition in these districts in particular.

[12] See previous footnote.

ensuing cognitive demand during the implementation of these tasks in classroom (Stein, Grover, & Henningsen, 1996; Stein & Lane, 1996). This series of rubrics for classroom observation align with the goals and three-part *launch-explore-summarize* structure of the inquiry- and discussion-based approaches to math teaching and learning adopted by our districts. While the entire IQA instrument consists of 20 scales which are intended to categorize the rigor of students' learning opportunities (Boston, 2012), for this study, I restrict my analysis to the three scales which most broadly characterize the academic rigor of classroom instruction and which correspond to the lesson format which teachers in our study were attempting to implement. The three rubrics utilized in our study are: *task potential, academic rigor of task implementation,* and *academic rigor of discussion* (see Appendix B).

In this framework, the task potential is described as the complexity or rigor of student thinking required of students to successfully complete the task as it appears in print form in curricular or instructional materials. In contrast, the task implementation rubric intends to characterize the rigor of what actually occurs in the classroom, asking the question: *At what level did the teacher guide students to engage with the task in implementation?* In practice, students may not consistently have opportunities to engage in the high levels of thinking called for by a rigorous task as it appears in the intended curriculum. For example, teachers have been observed lowering the cognitive demand of tasks by telling students to complete only part of the written task (Garrison, 2013; Jackson, Garrison, Wilson, Gibbons, & Shahan, 2013). Alternately, teachers also lower the cognitive demand of the task when they provide multiple, step-by-step examples illustrating the solution to a similar problem scenario (Garrison, 2013; Jackson, et al., 2013; Henningsen & Stein, 1997). When this occurs, students are no longer required to develop genuine problem-solving skills but are instead asked to apply a mathematical procedure which has been explicitly specified by the teacher. Additionally, students might not be able to realize the levels of rigorous thinking called for in a task when expectations are unclear, when the classroom environment is distracting or chaotic, or when tasks are not appropriate given students' current knowledge and skills (Henningsen & Stein, 1997; Stein & Lane, 1996).

Scores on the task potential and implementation rubrics range from 1 to 4, with a score of 1 representing student thinking that requires only the recall of memorized terms, definitions, or formulae. Scores of 2 are assigned when students apply prescribed mathematical procedures to

calculate answers for problems. Level 3 tasks require students to cultivate meaning around the application of a mathematical procedure and to make connections to underlying mathematical ideas in the task through identifying patterns, making conjectures, or using multiple problem-solving strategies or representations. Finally, scores of 4 are reserved for tasks and activities which have all of the qualities required for a score of 3 but which also explicitly require students to explain and justify their solution and method.

The discussion rigor rubric is scored on a 0 to 4 scale and guided by the key question: *During a whole-class discussion following work on the mathematical task, to what extent did students show their work and explain their thinking about the important mathematical content?* A score of 0 indicates there was no concluding whole-class discussion. A score of 1 indicates that students provide brief or one-word answers in a whole-class discussion. A score of 2 indicates that, in a whole-class format, students describe their written work for solving the task but do not engage in a discussion of their strategies, procedures, or mathematical ideas. A score of 3 indicates that students show or describe their written work for solving a task and/or engage in a discussion of the important mathematical ideas in the task. During a level 3 discussion, students provide explanations of why their strategy, idea, or procedure is valid and/or begin to make connections between procedures and mathematical concepts, but the explanations and connections provided are not complete and thorough. A score of 4 indicates that, during the discussion, students provide thorough explanations of why particular strategies are valid and make connections between these strategies and the underlying mathematical ideas.

In the early spring of each years of the MIST project, research team members video-recorded two (ideally consecutive) mathematics lessons conducted by each of the approximately 120 teachers participating in the study. Trained coders later scored these videos using the IQA rubrics. Each year, interrater reliability was established and monitored on an ongoing basis; across the three academic rigor rubrics and across the four years of data collection, percent agreement averaged 70.5%, with kappa scores averaging 0.50.[13] These reliability statistics are comparable to those from other classroom observation instruments used in the MET Project (see Table 4). Well-cited rules of thumb would characterize these reliabilities as ranging from "fair"

[13] Kappa scores are measures of reliability based on percent agreement but adjusted for the probability of chance agreement given the actual distribution of the data (J. Cohen, 1960).

to "substantial" (Landis & Koch, 1977), although Hartman, Barrios and Wood (2004) suggested that lower agreement rates (in the range of 70%) are to be expected of more complex instruments and can, in some circumstances, be considered sufficient.

Table 4. Percent agreement (kappa statistic in parentheses) for the three academic rigor rubrics of the Instructional Quality Assessment (IQA), by study year. Overall reliability measures compared with those from classroom observation instruments as reported data from the Measures of Effective Teaching (MET) Project (source Park, Chen, & Holtzman, 2014). Classroom observations instruments from the MET project include the Classroom Assessment Scoring System (CLASS), Framework for Teaching (FfT), Mathematical Quality of Instruction (MQI), and the Protocol for Language Arts Teaching Instruction (PLATO).

	Y1	**Y2**	**Y3**	**Y4**	**Avg. Y1-4**
Task Potential	59.4 (0.37)	56.9 (0.29)	75.0 (0.63)	59.1 (0.36)	62.6 (0.41)
Task Implementation	78.1 (0.51)	78.5 (0.37)	89.3 (0.75)	63.6 (0.29)	77.4 (0.48)
Discussion	78.1 (0.71)	69.2 (0.58)	67.9 (0.55)	70.5 (0.59)	71.4 (0.61)
Yearly Average	71.9 (0.53)	68.2 (0.41)	77.4 (0.64)	64.4 (0.41)	**70.5 (0.50)**
	MET CLASS	MET FfT	MET MQI	MET PLATO	**MIST IQA**
	34 (0.21)	57 (0.24)	76 (0.51)	59 (0.45)	**70.5 (0.50)**

Data Structure

The initial sample under consideration for this analysis included all teachers for whom there were both IQA observation data available and teacher-linked student-level achievement data with which to estimate teacher value-added. After reviewing literature on the measurement properties of this classroom rubric (Wilhelm & Kim, 2015) and reliability of classroom observation measures more generally (Kane & Staiger, 2012), I decided to limit the analytical sample to teachers-years cases where two IQA lesson scores are available. A single composite IQA score was estimated for each teacher-year using principal factor analysis (Preacher & MacCallum, 2003) of two sets of observation scores from the IQA task potential, task implementation, and discussion subscales. Because of the excessive noise introduced in teacher value-added estimates when relatively low numbers of students are associated with each teacher (Ballou & Springer, 2015), analysis was further restricted to teacher-year observations with thirty

or more associated student test scores. These restrictions diminished the analytical sample from 302 to 226 teacher-year cases in District B; the number of District D cases fell from 310 to 211 (see Table 5. Change in analytical sample…).

Table 5. Change in analytical sample of teacher-year cases resulting from additional restriction of number of IQA observations and number of teacher-linked student records.

District	Total Teacher-Year obs with IQA and value-added data	Restricted to two IQA obs per year	Restricted to >29 students contributing to value-added estimation
B	302	247	226
D	300	225	211

Over the course of seven-year of data collection for this research project, additional schools were recruited for participation beginning in Year 5 in both Districts B and D, with the number of participating teachers in each district approximately doubling. In District D in the beginning of Year 7, some participating schools discontinued their participation in the research project and other schools were recruited to replace them. In order to account for some of the bias which might be introduced through changes in the participating schools over time, each school was assigned a different cohort code corresponding to district membership and specific years of participation (see Table 6 Change in school participation…). This cohort code was used to include a cohort fixed-effect in regression analysis.

Table 6. Change in school participation in research project, Districts B & D, Years 1-7

School Code	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Cohort Code
B1	X	X	X	X	X	X	X	1
B2	X	X	X	X	X	X	X	1
B3	X	X	X	X	X	X	X	1
B4	X	X	X	X	X	X	X	1
B5	X	X	X	X	X	X	X	1
B6	X	X	X	X	X	X	X	1
B7	X	X	X	X	X	X	X	1
B10					X	X	X	2
B1B					X	X	X	2
B15					X	X	X	2
B17					X	X	X	2
B26					X	X	X	2
D1	X	X	X	X	X	X	X	3
D2	X	X	X	X	X	X	X	3
D3	X	X	X	X	X	X	X	3
D4	X	X	X	X	X	X	X	3
D5	X	X	X	X	X	X		4
D6	X	X	X	X	X	X	X	3
D7	X	X	X	X	X	X		4
D8					X	X	X	5
D13							X	6
D15					X	X	X	5
D17					X	X	X	5
D19					X	X	X	5
D20					X	X		7
D22							X	6
D24					X	X	X	5

CHAPTER IV

METHODS

In order to estimate the relationship between ambitious teaching in mathematics and teacher value-added estimates on different state tests, I employed the *average residual approach* to measuring teacher value-added. This approach has been utilized in a number of analyses using Measures of Effective Teaching (MET) Project data, which have examined the relationship between a number of different classroom observational measures and value-added scores across a variety of student assessments (e.g., Gates Foundation, 2010; Grossman, Cohen, Ronfeldt, Brown, 2014; Kane, McCaffrey, Miller, & Staiger, 2013; Polikoff, 2014; Polikoff & Porter, 2014; Ruzek, Hafen, Hamre, & Pianta, 2014; Walkington & Marder, 2014). Using the same estimation approach as these analyses will allow results from my analysis to be comparable to this relatively recent collection of research, allowing me to situate any findings in a greater body of evidence and provide for additional points of comparison.

In MET Project analyses, teacher-level scores were estimated using the following model (Gates Foundation, 2010; Kane et al., 2013):

$$Y_{it} = \beta X_{it} + \gamma \bar{X}_{jkt} + \theta Y_{it-1} + \lambda \bar{Y}_{jkt-1} + \varepsilon_{it} \quad [11]$$

where *Y* are scores on state assessments, standardized to have a mean of 0 and standard deviation of 1 within each state, year, and grade level. The *i* subscript represents the student, the *j* subscript represents the teacher, the *k* subscript represents the class or course section, and the *t* subscript represents the year. *X* is a vector of student level characteristics including race/ethnicity, gender, free- or reduced price lunch status, special education status, and English learner status; $\bar{X}_{jkt}$ represent the mean of these student characteristics at the class level, and $\bar{Y}_{jkt-1}$ represents mean prior achievement scores , also at the class level. With these models fitted separately for each district and grade level (Gates Foundation, 2010), year-specific teacher-

level value-added estimates, $\hat{\tau}_{jt}$ were generated by calculating averaging residuals at the teacher-year level ($\bar{\varepsilon}_{jt}$), as in the following equation:

$$\hat{\tau}_{jt} = \bar{\varepsilon}_{jt} \quad [12]$$

MET Project authors equate this approach to estimating teacher value-added as being equivalent to teacher fixed-effects approaches (Gates Foundation, 2010; Kane et al., 2013) and add that random-effects estimates in practice correlate very highly with teacher fixed-effects estimates, given that a large proportion of the variation in student outcomes is within classrooms and teachers. In subsequent published analyses using MET Project data, relationships between these estimates of teacher value-added and other teacher-level variables were investigated using correlation analysis (e.g. Ruzek et al., 2014; Polikoff 2014, Walkington & Marder, 2014) and/or regression analysis (Grossman et al., 2014; Kane et al., 2013; Ruzek et al., 2014; Polikoff & Porter, 2014).

Alternate Specifications for Robustness Checks

Various studies of teacher effectiveness have used different model specifications to estimate teacher value-added. The model producing the primary estimates of teacher effectiveness of interest for this analysis (as described above) was chosen in part because of its adoption in a large nation-wide study of measure of teacher effectiveness. However, researchers contest which models are the most appropriate for estimating teacher contributions to student learning, and many alternate methodological approaches present a reasonable options, each with its own advantages, disadvantages, and tradeoffs. For this reason, fourteen alternate specifications for estimating teacher value-added have been identified from the literature to serve in checking for robustness in this dissertation analysis (see Appendix C).[14] Robustness checks will employ these fourteen sets of estimates as outcome measures. These robustness analyses will follow the approach of the primary analysis described below, reporting results from three

[14] As noted previously in this dissertation, approaches to modeling and estimating teacher contributions to student learning can be sorted into three larger categories: (1) univariate response value-added models, (2) multivariate response value-added models, and (3) student growth percentiles. Both the main model and the robustness checks examined will only fall within the first category.

regression approaches: pooled ordinary least squares (POLS), hierarchical linear growth curve (HLGC) modeling, and teacher fixed effects (TFE).

Describing the Relationship between Characteristics of Instruction and Value-added Estimates

After estimating value-added estimates for teacher *j* in year *t* (Equation 12), I then estimate as a second stage pooled OLS estimate:

$$\hat{\tau}_{jt} = \pi_0 A_t + \pi_1 (A_t \times P_{jt}) + \varepsilon'_{jt} \quad [13]$$

where *P* is a measure of classroom practice for teacher *j* in year *t*. The interaction of these variables with *A*, an indicator variable for each assessment, will allow the relationship between instructional practices (in vector *P*) and the teacher-effectiveness estimate derived from assessment scores ($\hat{\tau}_{jt}$) to vary by assessment, allowing for a different estimated average relationship between teachers' IQA score and estimated value-added, depending on whether the year of observation corresponds to CCR aligned assessments (i.e. years 5 to 7 of the study) or if the observation occurred before CCR standards were adopted (i.e. years 1 to 4 of the study, *pre-CCR*).

As a second regression approach, teacher fixed-effects estimations contain a teacher-level fixed effect, π_j, as in Equation 14:

$$\hat{\tau}_{jt} = \pi'_0 A_t + \pi'_1 (A_t \times P_{jt}) + \pi_j + \varepsilon''_{jt} \quad [14]$$

For a third and final regression approach, a linear growth curve model is fitted:

$$\begin{aligned} \hat{\tau}_{jt} = (G_0 + g_j) + (B_0 + b_j) A_t + (C_0 + c_j) Time1_{t=\{1,2,3\}} \\ + (D_0 + d_j) Time2_{t=\{4,5,6\}} + E_0 P_{jt} \\ + F_0 (A_t \times P_{jt}) + \varepsilon_{jt} \end{aligned} \quad [15]$$

where *Time1* is a variable coded {0,1,2,3} for study years 1, 2, 3, and 4, respectively, and *Time2* is a variable coded {0,1,2} for study years 4, 5, and 6, respectively. These variables, together with the general intercept G_0 and allowance for a discontinuity or CCR average effect (B_0), can model two potentially distinct linear time trends for teacher value-added $\hat{\tau}$ in the Pre-CCR or

CCR years (i.e A_t=0 or A_t=1).[15] This model includes normally distributed random effects (g_j, b_j, c_j, and d_j) to model deviations of individual teacher growth trajectories from the group average trajectories. As in Equations 12 and 13, the primary variables of interest are the coefficients on the P_{jt} and ($A_t \times P_{jt}$), which describe to the extent to which observed instructional practices predict estimated teacher value-added in the pre-CCR time period, as well as the extent to which this relationship changes, if at all, during the CCR time period.

[15] With the exception of the random effects, this specification is akin to a regression discontinuity model

$$y_i = \alpha + \tau D_i + f(X_i) + I(D > 0) f(X_i) + \varepsilon_i$$

which allows for one slope (*f*) for values of *X* to the left of the cut-off value *D*, as well as a potentially different slope (*If*) for values of *X* greater than *D*, as well as allowing for a vertical displacement or discontinuity in the piecewise function for values of $X>D$ with the addition of the τD_i term. (Example taken from Doyle, Lee, & Nguyen, 2017)

CHAPTER V

RESULTS

Dependent Variable: Teacher value-added Estimates: Main Model

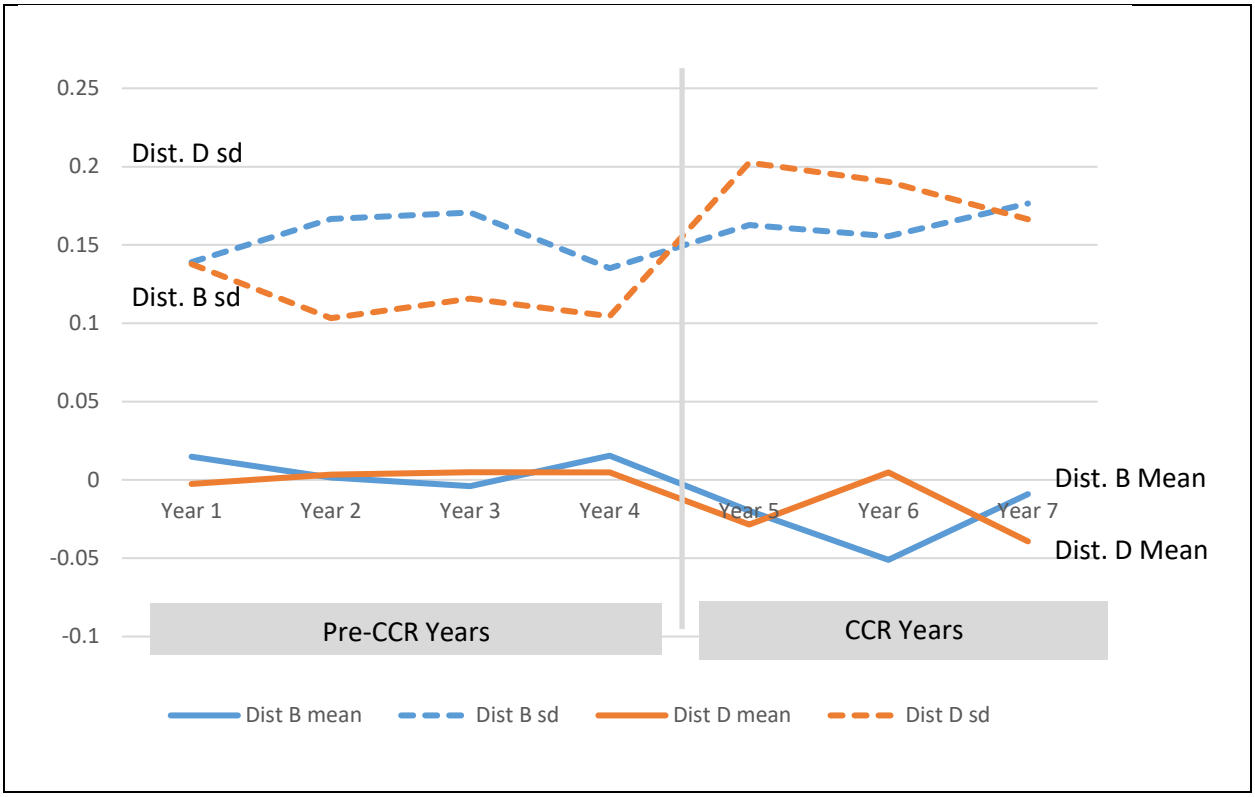
In every year and every district, the average estimated value-added from teachers included in this analytical sample is non-zero (Table 7, Figure 1). Teacher value-added was estimated each year using all of the available student-level data from all district schools participating in the study, even though not all teachers in all schools were participating in our study. In many-cases, students were assigned to teachers who were not full participants in the study, whose classroom interactions were not observed, and who are not included in the analytical sample. For that reason, when the average teacher value-added estimate deviates from zero in a given year and district, it should be interpreted as an indication that, on average, students in the classes of participating teachers tended to under- or outperform similar students in the classes of non-participating teachers. In other words, when average teacher-value added deviates from zero, it suggests that participating teachers, as a group, may have been more or less effective than non-participating teachers in the same schools, with effectiveness measured here by teacher value-added. While these mean-level differences between participating and non-participating teachers are non-zero, none are statistically significant at conventional levels.[16]

[16] As a point of comparison, Kane and Staiger (2012) report average district-level value-added estimates for their participating teacher to range from -0.012 to +0.020 across their six participating school districts in the 2009-2010 school year.

Table 7. Primary estimates of teacher value added descriptive statistics, by district, by pre/post CCR assessment.

	Dist B, pre-CCR	Dist B, CCR	Dist D, pre-CCR	Dist D, CCR
Variation across teachers	0.016	0.013	0.008	0.012
Variation within teachers across time	0.007	0.014	0.005	0.019
	Dist B, pre-CCR	Dist B, CCR	Dist D, pre-CCR	Dist D, CCR
Variation across teachers	70%	48%	61%	38%
Variation within teachers across time	30%	52%	39%	62%
Mean	0.006	-0.028	0.003	-0.019
Sd	0.153	0.164	0.114	0.183
N	106	120	99	112
Observations Per Teacher				
Teachers with 1 observation	27	54	37	63
Teachers with 2 observations	13	21	14	14
Teachers with 3 observations	11	8	6	7
Teachers with 4 observations	5	NA	4	NA
Total Teachers	**56**	**83**	**55**	**84**

Figure 1. Change in average teacher value-added estimates among full-participants in the analytical dataset, over time, Districts B & D.

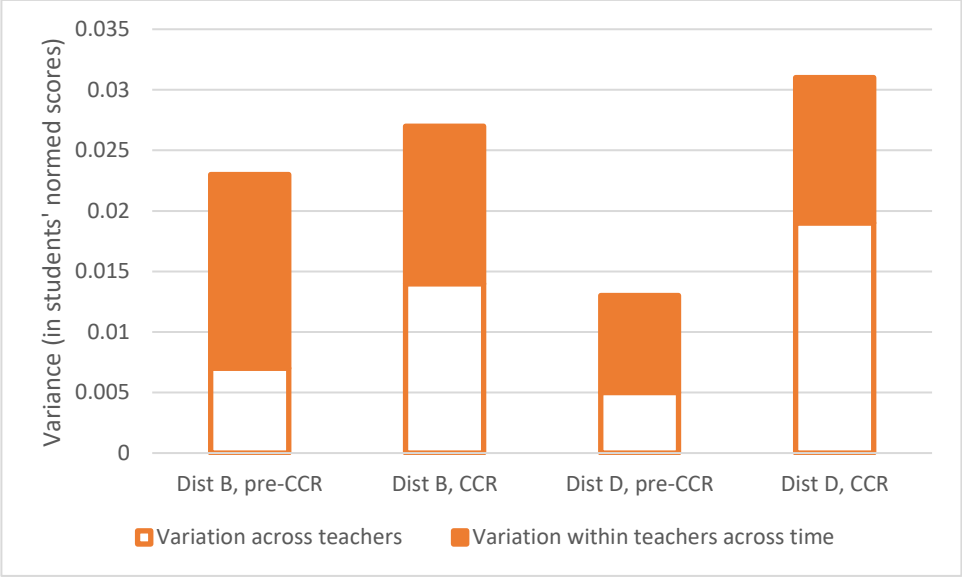


In both the pre-college- and career-ready assessment time period (*pre-CCR*, school years 2007-08 to 2010-11) and the college- and career-ready assessment time period (*CCR*, school years 2011-12 to 2013-14), the teacher value-added estimates produced through the primary estimation model are consistent with teacher effect sizes described in previous literature (e.g. Hanushek & Rivkin, 2010; Nye, Konstantopoulos, & Hedges, 2004). In District B in the pre-CCR time period, students with a teacher at the first standard deviation of effectiveness score an average of 0.15 standard deviations higher than similar students in participating schools. In the CCR time period in District B, the teacher effect size is estimated to be very similar in size, at 0.16 standard deviations (Table 7. Primary estimates of teacher…). In District D, the estimated teacher effect size is somewhat smaller in the pre-CCR time period (0.011 standard deviations), but then increases considerably in the CCR time period, to 0.18 standard deviations. To provide

some context, reviews of prior literature report estimated teacher effect sizes in math to fall between 0.11 and 0.36 standard deviations (Hanushek & Rivkin, 2010).

The within-teacher stability of these value-added estimates also falls within values reported in prior literature (Arronson, Barrow, & Sanders, 2007; Ballou, 2005; Goldhaber & Hansen, 2013; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009), with these scores being more stable in the CCR years than in the pre-CCR time period (Table 7, Figure 2). In District B, interclass correlations for teacher value-added scores were estimated to be 0.30 in pre-CCR years and 0.52 in the CCR time period. A larger increase in stability over time was noted in District D, where the interclass correlation is estimated at 0.39 in the pre-CCR years and 0.62 in the CCR time period. As a point of comparison, McCaffrey and colleagues (2009) utilized data from middle school students in Florida and find year-to-year correlations of value-added scores to vary between 0.3 and 0.6.

Figure 2. Partition of variance of teacher value-added, by district, by assessment.



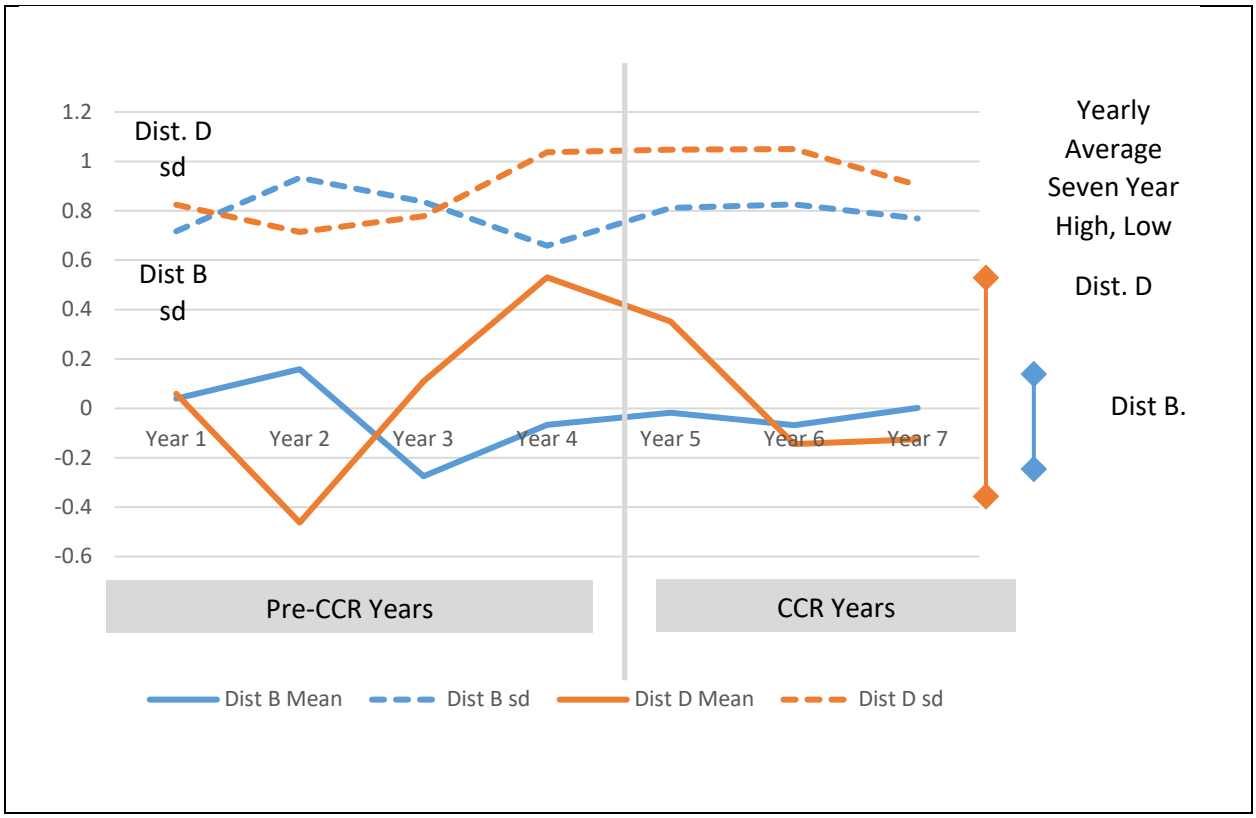
A number of scenarios could result in greater variation in estimated teacher effect. While variation over time of both (1) the spread of the distribution of teacher value-added estimates and (2) year-to-year variation in an individual teacher's scores may reflect real changes in between-teacher effectiveness or changes in individual teacher effectiveness over time, a number of other

factors may introduce this kind of increased estimation of teacher effect. Some of these factors include changes in class-size and teacher-student contact time, changes in the mechanism of peer effects, changes in ability tracking practices, and changes in student-level progress monitoring and intervention practices (Sass, 2008). However, increases in the variation of estimated teacher effect size could also be due to changes in the way student knowledge is assessed. Sass (2008) noted that changes in student achievement assessment properties, especially those close to the assessments' floor- and ceiling effects, could influence the size of teacher effect estimates. Additionally, if a new assessment measures different knowledge and skills, and the focal population of teachers has a greater variation in their effectiveness at teaching this newly assessed body of knowledge and skills (compared to their variation in teaching the previous body of knowledge and skills), then increased dispersion of teacher effect might be expected, as in these data.

Descriptive Statistics: IQA Composite

In general, average IQA scores in District B could be characterized as showing a less distinct time trend and smaller year-to-year deviations from the seven-year average than comparable data from District D (Figure 3). District D data display more distinct time trends, with average IQA composite scores lowest in at -0.46 in Year 2, rising to a peak of + 0.53 in Year 4, and declining in each subsequent year. While the IQA composite yearly averages in District B deviate from the district's seven-year average by only plus or minus 0.25 standard deviations, year-to-year deviations in District D are as large as plus or minus 0.51 standard deviations from the 7 year average.

Figure 3. Change in average IQA composite score over time, Districts B & D



A partition of variance analysis of the IQA composite over time also suggest that individual teachers' quality of classroom instruction, as measured by the IQA, is more stable in District B than in District D (Figure 4, Table 8). In District B, the total variation in IQA composite scores changed little from the pre-CCR time period to the CCR time period. In addition, the ratio of the variation in within-to-between teacher variation in scores does not seem to change substantially between these two periods in District B. In contrast, in District D, there is more variation in IQA overall than in District B. This overall variation in IQA scores increases in District D in the CCR years, and the estimated proportion of the variation which is between teachers in this time period is negligible compared to the whole (i.e. not distinguishable from zero).

Figure 4. Decomposition of variance of IQA composite score, by district by test regime

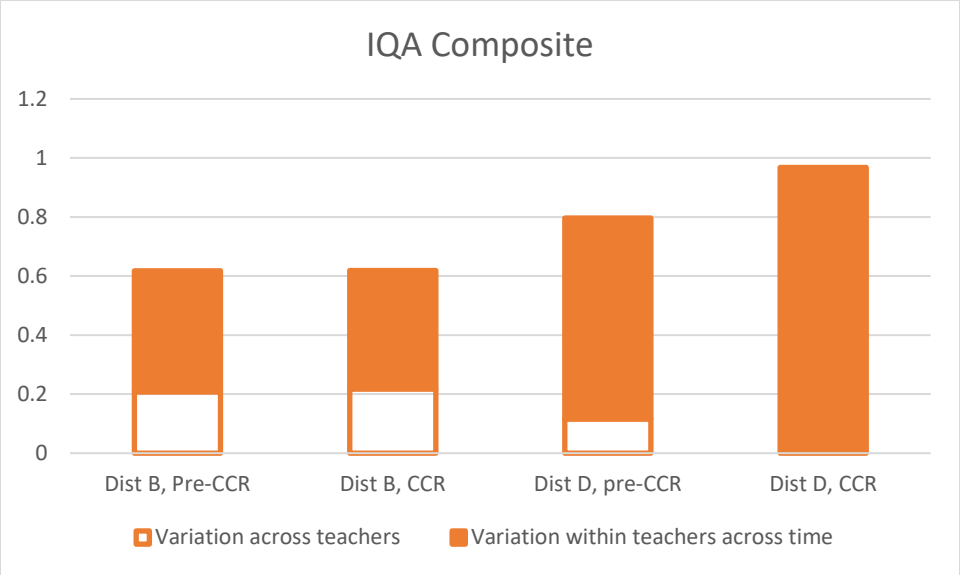


Table 8. Decomposition of variance of IQA composite score, by district by test regime.

	Dist B, Pre-CCR	Dist B, CCR	Dist D, pre-CCR	Dist D, CCR
Variation across teachers	0.210	0.233	0.149	0.000
Variation within teachers across time	0.422	0.412	0.649	1.000
	Dist B, Pre-CCR	Dist B, CCR	Dist D, pre-CCR	Dist D, CCR
Variation across teachers	33%	36%	19%	0%
Variation within teachers across time	67%	64%	81%	100%
Mean	-0.050	-0.029	0.050	-0.044
SD	0.801	0.797	0.904	1.00
N	106	120	99	112

Analysis Results: Correlations: Comparisons with MET Data

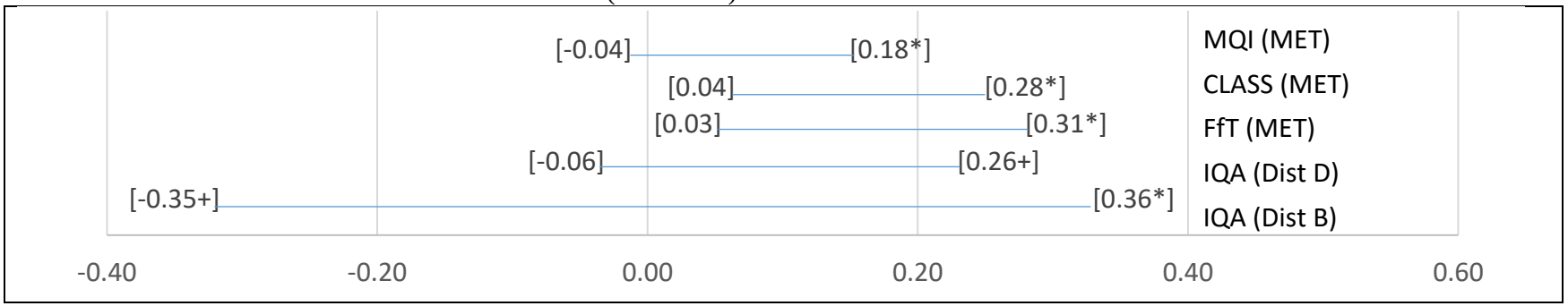
As stated previously, some of the motivation in choosing the approach to estimating teacher value-added utilized by the MET Project was to compare findings using this dataset with findings from that study. Using data from one school year, Polikoff (2014) demonstrated some variation in the correlations between observational scores and value-added estimates across districts. Polikoff also reported an "overall" correlation: the correlation coefficient calculated after pooling these observations across districts, with each observation given equal weight (for detail, see Appendix D). Here, I attempt to use a similar table to explore the data from the current study and show how Polikoff's correlations of the MET Project's measures across districts in a single year compares to correlations of the IQA and value-added within each individual MIST district across several years (see Table 9).

The average IQA to value-added correlations from the District D data are similar to those the correlations between mathematics classroom observational measures and estimated value-added correlations reported by Polikoff (2014) – generally between 0.10 and 0.20. Furthermore, the year-to-year variation in District D correlations was similar to the between-district variation in data reported by Polikoff. While the District D correlations are comparable to those reported in the MET Project data, the District B correlations contrasted with these data. While some MET project district-by-year correlations were negative in sign for a given district, in no instance of their data was the average correlation across districts less than zero. Not only is the average value-added to classroom observation correlation across years negative for District B (-0.10), but there was also much larger variation in the correlations calculated each year (see Appendix D for detail). For example, the IQA composite-value added correlation in District B was -0.35 ($p<0.10$) in Year 1, and then -0.36 ($p<0.05$) in Year 4. The greater variability in year-to-year correlations in District B is also noted in that the standard deviation of the collection of correlation coefficients is much larger in District B (0.22) than it is for District D or the MET Project data ($\geq 0.09, \leq 0.12$).

Table 9. Average correlations between classroom observation scores and value-added in mathematics estimates using the MET Project methodology. Correlations shown here (corresponding to Framework for Effective Teaching (FFT); Classroom Assessment Scoring System (CLASS), and the Mathematical Quality of Instruction (MQI) rubric) are averaged across districts. Correlations reported from the current study (IQA composite) are averaged within district across years. (See Appendix D for detail).

Classroom Observation Rubric	Average Correlation with value-added[17]	Corr. min	Corr. max	Districts	Years	Teacher-Years	Time Period (School Years)
FFT	0.19	0.03	0.31*	5 (Districts #1-2,4-6)	1	805	09-10
CLASS	0.16	0.04	0.28*	5 (Districts #1-2,4-6)	1	804	09-10
MQI	0.04	-0.04	0.18*	5 (Districts #1-2,4-6)	1	794	09-10
IQA	-0.10	-0.35+	0.36*	1 (District B)	7	226	07-08 to 14-15
IQA	0.13	-0.06	0.26+	1 (District D)	7	211	07-08 to 14-15

Figure 5. Comparison of inter-district range of observation score –value-added correlation coefficients (MET data) with intra-district range of observation score –value-added correlation coefficients (MIST data).



[17] Polikoff (2014) reported the individual correlation coefficients between classroom observation measures and estimated teacher value-added in five different districts, as well as an "overall" correlation which pools observations from the five districts. In his measure of "overall" correlation, each observation appears to be weighted equally, regardless of the population or sample size from each district. However, interpretations of correlation coefficients and associated p-values becomes problematic when observations are not independent (as in the case of the MET data where teachers may be drawn from the same school or district, or in the case of the data from the current study, where observations in consecutive years may be drawn from the same teacher) (Havlicek & Peterson, 1977). With these limitations in mind and purely for descriptive and comparative purposes, I report here the average (i.e., arithmetic mean) of the correlations across districts (in the case of data from MET/Polikoff 2014) or across years (in the case of data from the current study), and opt not to attempt to calculate or report p-values.

Because the correlation between the IQA composite and value-added showed substantial variation within district from year to year, and because of the relatively small sample sizes in this district-by-year analyses, the District B and D data were analyzed for the presence of outliers with substantial influence on the regression coefficients. Using a DFBETA procedure (Besley, Kuh, & Welsch, 1985; Bollen & Jackson, 1990), each teacher-year observation was scrutinized for its influence. Although Belsley, Kuh, and Welsch (1985) suggest scrutiny of observations with an absolute DFBETA value of larger than $2/\sqrt{n}$, in order to retain more observations, a more lenient cutoff suggested by Bollen and Jackson (1990) was employed, identifying observations with absolute DFBETA values larger than 1.0. Using this procedure and cutoffs, three high-influence observations were observed and removed from the analytical data set. Removing these three observations resulted in a stronger overall correlation in District D data (with the correlation coefficient average increasing from 0.13 to 0.16) and a weaker but still negative average correlation in District B (-0.10 to -0.05, see Table 10).

Table 10. Analytical sample after omission of high-influence outliers.

Classroom Observation Rubric	Average Correlation with value-added[18] (unweighted)	Districts	Years	Teacher-Years	Time Period (School Years)
IQA	-0.05	1 (Dist. B)	7	224	07-08 to 14-15
IQA	0.16	1 (Dist. D)	7	210	07-08 to 14-15

This correlational approach to analyzing these data has substantial limitations. The nested nature of these data is not accounted for in this approach to correlational analysis, violating assumptions necessary for making accurate inferences from these correlation coefficients (Havlicek & Peterson, 1977). For that reason, the regression analysis in the next section is expected to allow for more informative and robust inferences.

[18] See Footnote 15

Analysis Results: Regressions

District B

Results from District B data generated no statistically significant findings with regards to the primary research question: *Compared to pre-CCR tests, do CCR era tests yield teacher value-added estimates that are more sensitive to ambitious math instruction*? Across estimation approaches, the regression coefficient for IQA composite variable in pre-CCR ranged from +0.007 to +0.026, suggesting a slightly positive relationship with estimated value-added in pre-CCR years, although these estimates are small and not statistically significant at conventional levels (Table 11, Table 12). The IQA-composite-CCR test interaction term for District B was negative (-0.025 to -0.050), suggesting that in the CCR test years, higher IQA composite was associated with lower value-added in the CCR years than it was in the CCR years. However, this point estimate is also not statistically significant at conventional levels. Because the sum of the IQA coefficient and the CCR interaction term is negative (-0.019 to -0.027), there is some suggestion that, in CCR years, higher IQA scores are associated with *lower* teacher value-added, although this relationship is not statistically significant at conventional levels (Table 11, Figure 6).

Table 11. Linear combinations of coefficients of teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. (Fully interacted mode). (P-values in parentheses).

	Fixed Effects	HLGC	POLS
IQA Composite in pre-CCR Years (Dist B)	0.026 (0.444)	0.015 (0.372)	0.007 (0.705)
IQA Composite in CCR Years (Dist B)	-0.025 (0.282)	-0.022 (0.372)	-0.017 (0.330)
IQA Composite in pre-CCR Years (Dist D)	0.003 (0.814)	0.013 (0.391)	0.011 (0.346)
IQA Composite in CCR Years (Dist D)	0.020 (0.268)	**0.034*** (0.022)	**0.048***** (0.000)
Difference :IQA Coef. in Dist. D and B (Pre –CCR Years)	-0.022 (0.539)	-0.002 (0.935)	0.004 (0.874)
Difference :IQA Coef. in Dist. D and B (CCR Years)	0.045 (0.127)	**0.057*** (0.015)	**0.066**** (0.003)
Difference: Pre-CCR to CCR Dist B	-0.050 (0.211)	-0.037 (0.120)	-0.025 (0.343)
Difference: Pre-CCR to CCR Dist D	0.016 (0.397)	0.021 (0.311)	**0.037*** (0.032)
Difference :IQA "slopes". in Dist D and B (ΔD-ΔB)	0.067 (0.135)	**0.058+** (0.066)	**0.062*** (0.048)

Table 12. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. (P-values in parentheses). (Note: Difference between IQA-Value-added in District D and B estimated using a single fully interacted model for both districts in all years (see Appendix E)).

	Dist B			Dist D		
	Fixed Effect	HLGC	POLS	Fixed Effect	HLGC	POLS
	(1)	(2)	(3)	(4)	(5)	(6)
IQA Composite (2 Day)	0.026	0.015	0.007	0.003	0.012	0.011
	(0.445)	(0.408)	(0.705)	(0.815)	(0.334)	(0.349)
IQA x CCR	-0.050	-0.034	-0.025	0.016	0.019	**0.037***
	(0.212)	(0.182)	(0.343)	(0.400)	(0.317)	(0.033)
CCR	-0.042	-0.025	-0.019	-0.075*	-0.042	-0.032
	(0.113)	(0.489)	(0.338)	(0.042)	(0.287)	(0.278)
Year Slope (pre CCR)		0.004			-0.002	
		(0.783)			(0.890)	
Year Slope (CCR)		0.012			-0.007	
		(0.448)			(0.724)	
Cohort Controls	X	X	X	X	X	X
Intercept	-0.007	-0.004	0.006	0.017	0.017	0.005
	(0.529)	(0.875)	(0.733)	(0.159)	(0.480)	(0.798)
N	224	224	224	210	210	210
AIC	-483.6	-204.9	-186.8	-537.3	-236.7	-193.6
BIC	-473.3	-181.1	-169.8	-523.9	-203.2	-166.8
IQA Composite in CCR Years (IQA coefficient + CCR interaction Term)	-0.025	-0.022	-0.017	0.020	**0.034***	**0.048*****
	(0.282)	(0.372)	(0.330)	(0.268)	(0.022)	(0.000)
Difference: IQA Coef. in Dist. D and B (Pre –CCR Years)				-0.022	-0.002	0.004
				(0.539)	(0.935)	(0.874)
Difference: IQA Coef. in Dist. D and B (CCR Years)				0.045	**0.057***	**0.066****
				(0.127)	(0.015)	(0.003)

Figure 6a-c. Comparison of point estimates by time period, by district, by estimation method (Results from fully-interacted regression – see Appendix D)

Figure 6a Teacher-Fixed Effects model

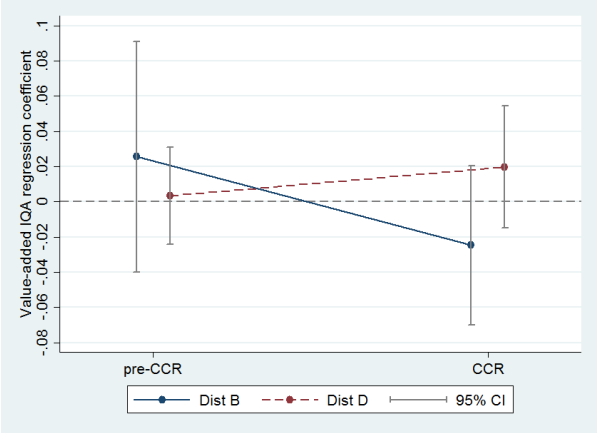


Figure 6b. Hierarchical Linear Growth Curve model

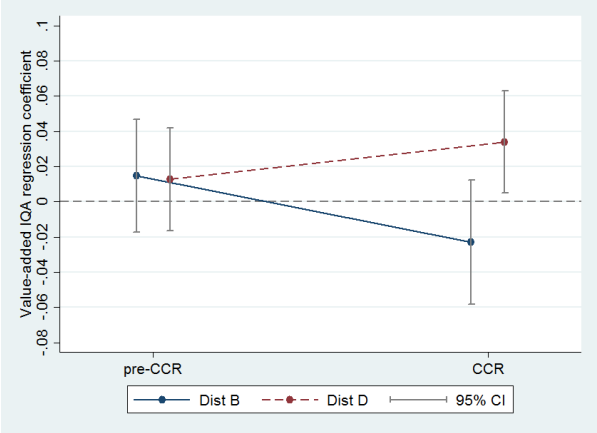
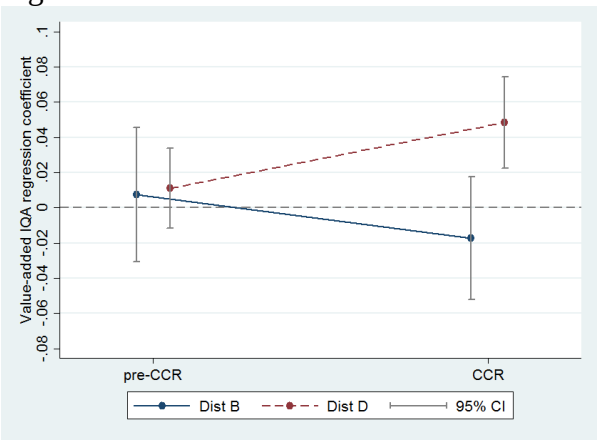


Figure 6b. Pooled-OLS model



District D

While most of the coefficients from the regression analysis were not statistically significant at conventional levels, one set of coefficients were notably statistically significant across more than one estimation method. The sum of the IQA coefficient and the CCR-IQA interaction term was positive in all three of the estimation methods, and significant at conventional levels in two of the estimation methods. The size of the coefficients in the HLGC and POLS models suggest that, in CCR years, a one standard deviation increase in the IQA composite scores was associated with that teacher's students performing 0.034 to 0.048 standard deviations greater on the end-of-year test, when compared to similar students in their cohort.

Using a common rescaling approach[19], this size gain is estimated to be equivalent to an additional 19 to 29 days of instruction (Table 11, Figure 6). In prior years, teachers' observed IQA composite scores were not associated with teacher value-added scores at conventional levels of significance, whereas there is some evidence to suggest that during the CCR years, these two measures were positively associated with each other at conventional levels of significance.

However, in answering the primary research question -- *Did the assessments become more sensitive to ambitious instruction during the CCR years* – the regression results from District D are less definitive. The IQA-CCR interaction term coefficients – which measure the change in the relationship between IQA scores and value-added in the CCR test years compared to the pre-CCR test years – are all positive in sign, but only one is significant at the 0.05 level. In the pooled ordinary least squares (POLS) model, the IQA-CCR interaction term is +0.037 (p=0.03), providing some evidence that the relationship between IQA and value-added is greater in the CCR years than in the pre-CCR years.

The data and analysis presented address the following research question: *Compared to the NCLB-era assessments they replaced, are "college- and career-ready" assessments of middle school mathematics more sensitive to ambitious teaching and learning of mathematics observed in the classroom? Compared to the assessments they replaced, is the relationship between the classroom observation of ambitious math instruction and teacher value-added stronger?* Drawing on the data and analysis here, we cannot say definitively whether or not the relationship between ambitious teaching and learning of mathematics and estimated teacher value-added was different after the end-of-year tests changed in either of the two districts represented in these data. These data and the analysis does not allow us to say that those changes were significantly different from zero. However, the signs of the interaction coefficients *suggest* that the relationship between the IQA composite and value-added did change in different ways in both districts with the introduction of new assessments. IQA regression coefficients from District B in both the pre-CCR and CCR assessment time periods (Table 11, Figure 6) suggests that after

[19] Specifically, Hattie (2008) equated 0.25 standard deviations of growth along a normalized distribution of test scores as equivalent to one year (or approximately 180 days) of instruction. This conversion rate is also used in Kane and Staiger, 2012. However, I use a slightly more conservative conversion of 0.31 standard deviations per 180 days of instruction, based on data specifically from middle-school mathematics assessments, as described in Hill, Bloom, Black, & Lipsey, 2008.

the switch in assessments, higher IQA was not associated with higher value-added estimates. District D data shows a different pattern: in the pre-CCR era, there is no statistically significant evidence that IQA scores are linked to higher value-added, while after the test switched, there is some statistically significant evidence supporting the hypothesis that higher IQA scores are linked to higher value-added (Table 11, Figure 6).

Robustness Check Results

In addition to value-added estimates which were produced using the MET Project methodology and utilized in the regression analysis described above, fourteen alternate value-add estimates were generated using specification options explored in the research literature (for example, Henry, Rose, & Lauren, 2014; Guarino, Reckase, & Wooldridge, 2014) (see Appendix C). For each set of these fourteen alternative teacher value-added estimates, the relationship between IQA composite score and value-added was examined using three approaches: teacher fixed effects (TFE), hierarchical growth curve modeling (HLGC), and pooled ordinary least squares (POLS).

Figure 7a-b. IQA-CCR Test interaction term coefficients from (a) main model value-added estimates and (b) robustness check regressions using alternate value-added estimations, plotted by coefficient value and p-value; District B

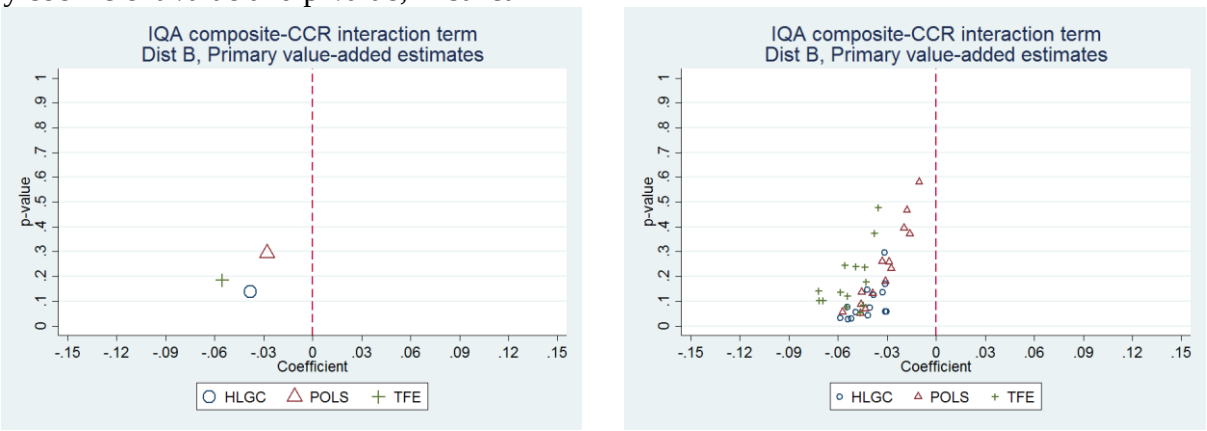
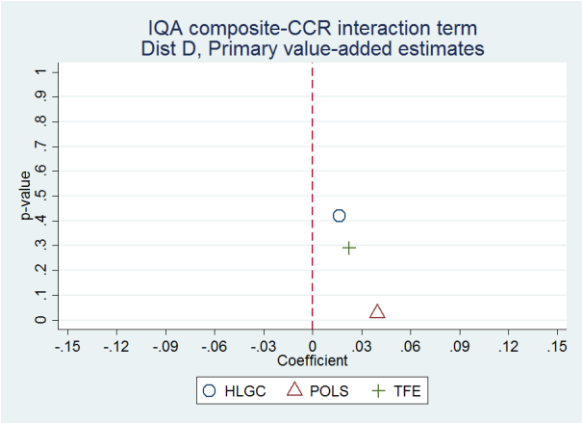
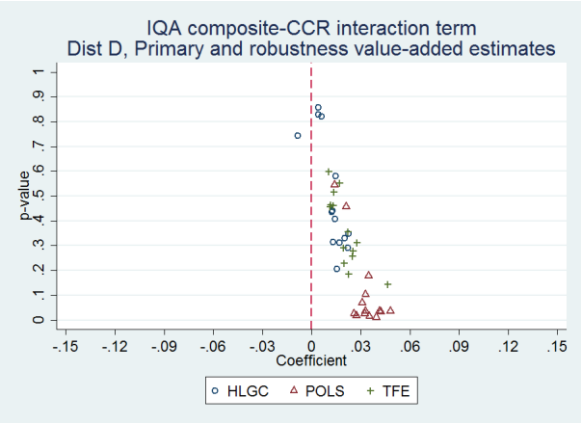


Figure 8a-b. IQA-CCR Test interaction term coefficients from (a) main model value-added estimates and (b) robustness check regressions using alternate value-added estimations, plotted by coefficient value and p-value; District D





The pattern of coefficients produced through regressions using these alternate sets of value-added estimates can provide some insight one two key issues: (1) the degree to which utilization of alternative value-added estimates might result in interaction term coefficients with drastically different signs or magnitudes, and (2) the extent to which choosing to utilize MET Project-style value-added estimates resulted in IQA-CCR test interaction terms which are at the very high or very low end of the distribution of coefficient estimates which would be produced were other value-added methodologies selected.

In looking at whether changing the value-added estimation approach would have changed the sign of the IQA-CCR test interaction term, we see that in the case of District B, *all* of the interaction term coefficients from these robustness check regressions are negative (Figure 7). In contrast, almost all of the interaction term coefficients in the District D analysis are positive (Figure 8). In conclusion, evidence presented here suggests that the sign of the interaction term coefficient is not sensitive to value-added specification.

Furthermore, these coefficients from the interaction terms which result from utilizing MET Project-style value-added estimates are not outliers when compared to interaction term coefficients resulting from regressions on alternatively-specified value-added estimates. In a distribution of fifteen coefficients (one from the MET Project approach to estimating value-added and fourteen using alternative value-added approaches), the coefficient values utilizing the MET Project approach are in the middle of the pack (in the 33rd to 73rd percentile) with none of

the six district-specific interaction term coefficients as an outlier (Table 13). This provides some evidence that, compared to regression results utilizing value-added estimate produced by the most commonly specified univariate response models for estimating value-added, regressions utilizing value-added estimates using the MET methodology do not produce extreme coefficient estimates in these data.

Table 13. Percentile rank of CCR interaction coefficient from main model, within distribution of coefficients from all value-added estimates.

District	Est method	Main model interaction term	For 14 alternate specifications			Pctl. rank of main model interaction term among alternate estimates
			Mean	Min	Max	
Dist B	TFE	-0.055	-0.052	-0.072	-0.035	33%
Dist B	HLGC	-0.038	-0.026	-0.061	-0.013	53%
Dist B	POLS	-0.028	-0.033	-0.057	-0.010	60%
Dist D	TFE	0.029	0.024	0.014	0.039	53%
Dist D	HLGC	0.014	0.023	-0.021	0.059	67%
Dist D	POLS	0.037	0.030	0.010	0.043	73%

CHAPTER VI

DISCUSSION

Review of Findings and Interpretation

My analysis of a longitudinal data set with scores from more than 39,000 student-year observations, nested in 244 teachers in 27 schools in two districts across seven years suggests that in one district (District D), prior to the adoption of CCR standards and assessments, the relationship between classroom measures of ambitious math instruction and student test score gains were not statistically distinguishable from zero at conventional levels of significance. However, in the three years subsequent to the move to CCR standards and assessments, more ambitious mathematics instruction in the classroom was associated with positive and statistically significant gains in student test scores in this district. Thus, there is some evidence to support that in this district in these grades and years, the test was sensitive to ambitious math instruction. Before the transition to CCR standards, the relationship between classroom interactions associated with problem-solving and higher-order thinking (as measured by the IQA classroom observation instrument) and increases in student learning were estimated to be similar in both districts: slightly positive (depending on the model used for estimation), but statistically indistinguishable from zero. However, with the transition to CCR standards, we see this relationship diverging. In the case of District D, the relationship in the CCR years between the classroom measures and value-added is estimated to be positive and significant at conventional levels in two of the three different regression approaches chosen for this analysis, with growth in this relationship from the pre-CCR years estimated to be positive and significantly different from zero in one of three models (Table 11, Table 12, Figure 6).

In District B, by contrast, the relationship in the CCR years between the classroom measures and value-added did not change from the CCR years. In fact, the regression results suggest that our best guess is that, in the CCR years, the relationship between the classroom observational measure and teacher value-added is actually *negative*, although both the point estimate in the CCR years and the estimated change in the relationship from the pre-CCR years to the CCR years are not statistically different from zero at conventional levels.

There are a number of reasons why, after a realignment of the intended and assessed curriculum to new college and career ready standards, we might expect to see teaching emphasizing rigorous, higher-order thinking associated with greater gains on students' assessment. As described in the review of literature at the beginning of this dissertation manuscript, this research question was situated in the logic of accountability and alignment as a vehicle for education reform: that a system of curricular tools, professional development, and assessments aligned to rigorous standards and accountability structures should be able to influence and improve the quality of classroom instruction and student learning (Hamilton, Stecher, & Yuan, 2008; O'Day & Smith, 1993; Smith & O'Day, 1991). This theory of change is predicated in part on assessments which can motivate more rigorous teaching and learning through accurately measuring higher-order thinking skills and then (1) holding students accountable for demonstrating this kind of learning and/or (2) holding teachers accountable for facilitating this kind of learning. In this approach to education reform, students in classrooms where rigorous and ambitious goals for learning mathematics are emphasized should, all thing being equal, receive higher scores on assessments, assuming that these assessments have been successfully designed to support these systemic goals.

In the case of District D, the data and analysis presented here suggest that the District D tests became more aligned to more ambitious goals for teaching and learning middle-school mathematics in this sample. The newer CCR tests resulted in teacher-level value-added estimates with a larger, positive, and statistically significant relationship with the IQA measure of rigor of classroom mathematics instruction. The test used in previous years resulted in teacher-level value-added estimates which have relationships with IQA which are generally positive, but smaller and not statistically significant. The size of the coefficients in the HLGC and POLS models suggest that, in CCR years, a one standard deviation increase in the IQA composite scores was associated with that teacher's students performing 0.034 to 0.048 standard deviations greater on the end-of-year test, when compared to similar students in their cohort. Using a common rescaling approach[20], this size gain is estimated to be equivalent to an

[20] Specifically, Hattie (2008) equated 0.25 standard deviations of growth along a normalized distribution of test scores as equivalent to one year (or approximately 180 days) of instruction. This conversion rate is also used in Kane and Staiger, 2012. However, I use a slightly more conservative conversion of 0.31 standard deviations per 180 days of instruction, based on data specifically from middle-school mathematics assessments, as described in Hill, Bloom, Black, & Lipsey, 2008.

additional 19 to 29 days of instruction. Given that the standard deviation of teacher value-added in District D during the CCR years is 0.183, it could also be estimated that a teacher who moves from average to one standard deviation above average on the IQA composite would progress positively on the distribution of teacher effectiveness by 0.19 to 0.26 standard deviations, equivalent to moving from the 50th percentile in teaching effectiveness to the 57th or 60th percentile.

There is some evidence that the relationships between student test score gains in middle school mathematics and composite IQA scores in District D in the CCR years are of similar magnitude to those of a number of other observational rubrics used in the MET Study (Kane & Staiger, 2012). While Kane and Staiger (2012) do not specifically report regression coefficients or p-values to describe the relationships between student gains on state assessments and classroom observation scores in their analysis of MET Study data, they do utilize a "running regression-line smoother" (p. 7) to describe a non-linear relationship between a teacher's percentile rank on a number of observation rubrics and the corresponding estimated gains on student assessments. The analysis in this dissertation describes a linear relationship between a standardized score on the IQA rubric and estimated gains on student assessments; however, these results can be rescaled to allow for comparison with Kane and Staiger's (2012) graphical analysis. A visual comparison of these relationships on equivalent scales suggests that the magnitude of the estimated relationship between the IQA composite and teacher value-added in District D during the CCR years is generally comparable in magnitude to the relationships associated with the MET study's observational rubrics (Figure 9a-b).

Figure 9a-b. Comparison of graphs of (a) teacher observation scores from Measures of Effective Teaching (MET) Study (Source: Kane & Staiger, 2012) with (b) predicted values using coefficients from regression results from District D in CCR years

Figure 9a

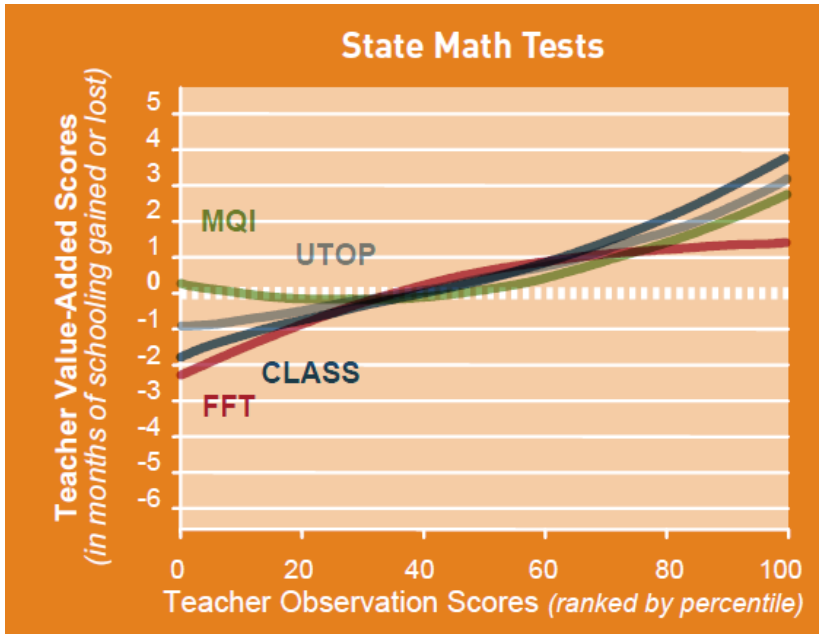
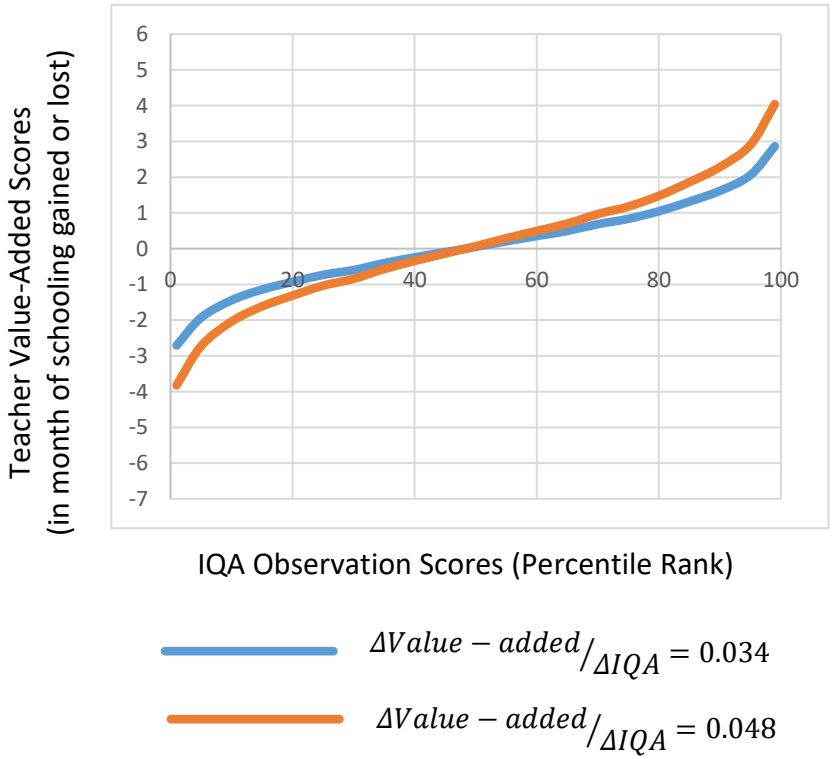


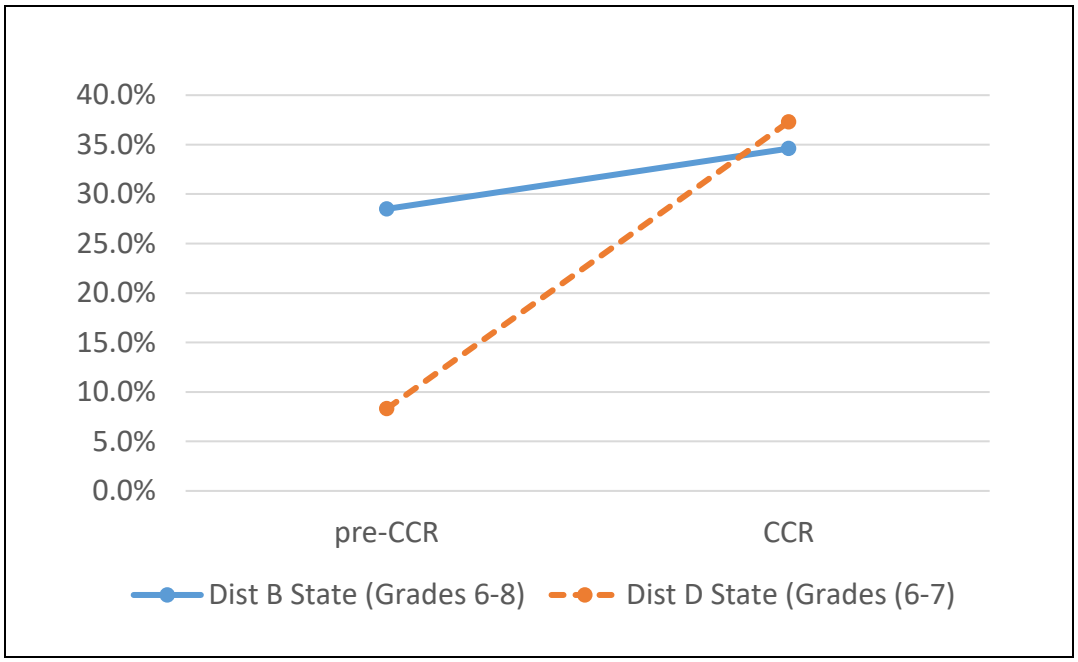
Figure 9b



The size of this relationship in District D in the CCR years can be compared to other effect sizes in the research literature focusing on student achievement in middle school mathematics. For example, the District D CCR year effect size is similar in magnitude to the relationship between student test score gains in middle school mathematics and mentor teachers' evaluations of mentee teachers (Rockoff & Speroni, 2011). The estimated improvement in gains associated with a one standard deviation increase in IQA in District D in the CCR years is also within the range of increases in middle school students' growth in state math assessment scores associated with assignment to teachers with one to three years of experience, compared to teachers with no prior experience (Chingos & Peterson, 2010; Harris & Sass, 2011; Ladd & Sorenson, 2017; Rice, 2010). Similarly sized increases learning gains in middle school mathematics are associated with assignment to a National Board Certified teacher (Chingos & Peterson, 2010; Cowan & Goldhaber, 2016), However, point estimates of IQA effect size of students' middle school score growth is smaller than 0.06, which is the estimated impact on that student outcome of being assigned to a Teacher for America Teacher (Clark et al., 2013) or one year of assignment to a KIPP school (Tuttle et al., 2015).

While the regression analysis presented here suggests improved alignment between ambitious instruction in mathematics and state assessments in District D but not in District B, other sources of evidence may both reinforce and explain the existence of this change in relationship.Sharpe, Rosenquist, and Kern (n.d.) analyzed the rigor of released items from middle school math state assessments from these two districts, looking for differences in item rigor between tests in pre-CCR years and those after the shift to college and career ready standards. In District D, the percentage of points possible associated with higher rigor items increased sharply in the CCR assessments of middle school math, compared to the pre-CCR assessments (see Figure 10).

Figure 10. Percent of total assessment points requiring high-rigor thinking, over time, by state/district (From Sharpe, Rosenquist, & Kern, n.d.)



This analysis also suggests that assessments in District B were more rigorous than those of District D during the pre-CCR years, and that District B's assessments saw smaller relative increases in rigorous items after the shift to CCR standards. We can then look for factors beyond test rigor in attempt to explain change – or lack of change – in the relationship in District B between ambitious instructions in mathematics and value-added derived from student test measures.

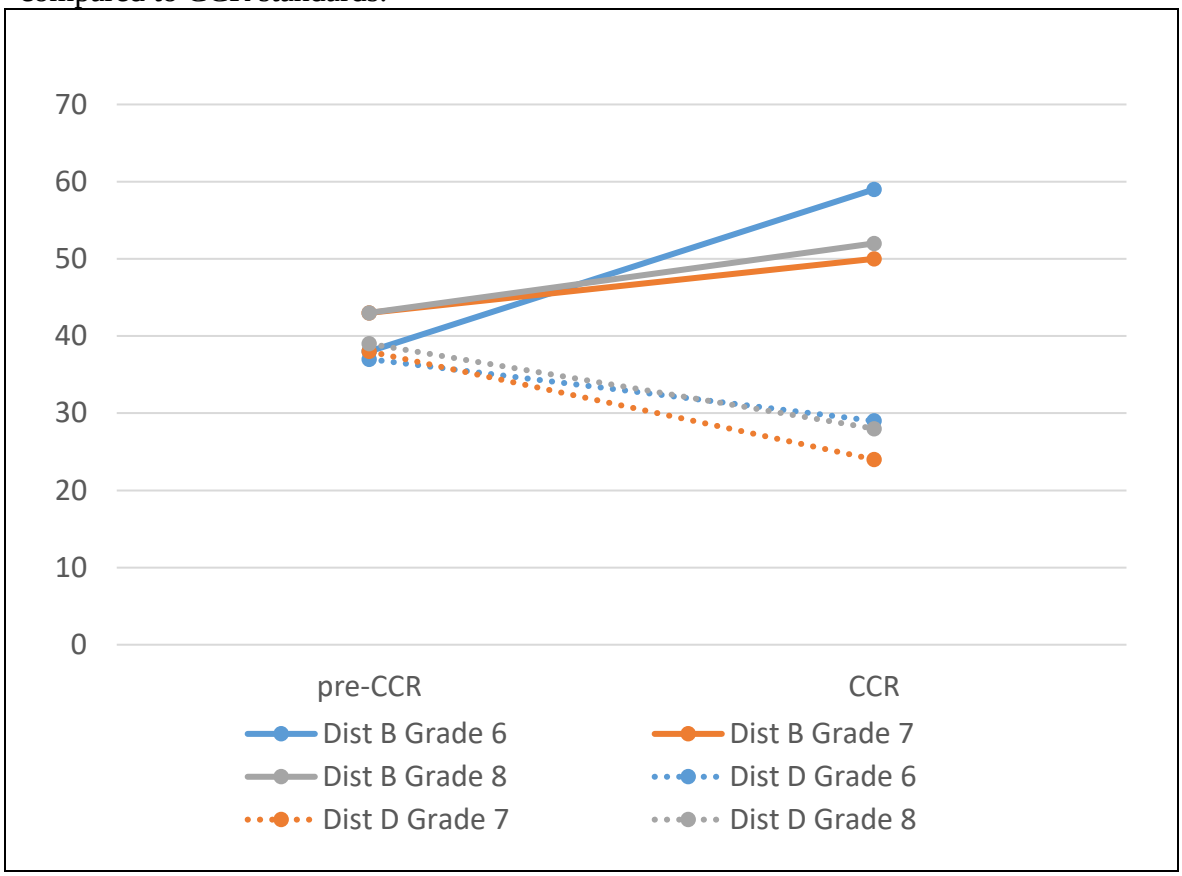
Most state-adopted CCR standards have embodied a purposeful shift to explicit expectations that students would study fewer topics, but study them in greater depth. In general, the shift to CCR standards has been described in the research literature as an intentional move towards greater focus on fewer learning standards, especially in mathematics, with the intent that students should learn fewer standards more thoroughly, as opposed to demonstrating a more superficial knowledge of more mathematical topics (Conley, 2014; Porter, McMaken, Hwang, & Yang, 2011). To this end, the CCR standards have been described as an improvement upon previous editions of state standards, which have generally been characterized as "a mile wide and

an inch deep" (Schmidt, McKnight, & Raizen, 1997). However, analysis of the shift in standards in District B indicate that in that state, the adoption of CCR standards resulted in more math standards rather than fewer: In middle school grades, the number of state standards in mathematics increased by 48%, 22%, and 4% in Grades 6, 7, and 8, respectively (Sharpe, Rosenquist, & Kern, n.d, see Table 14). In contrast, the state standards in District D became fewer and ostensibly more focused, with the number of standards in these grade *decreasing* by 22%, 37%, and 28% in Grades 6, 7, and 8, respectively (see Table 14, Figure 11, or Appendix I). After the shift in standards in both districts, District B documents list almost twice as many standards for middle grades math as does District D.

Table 14. Change in number of mathematical standards per grade, District B & D (source: Sharpe, Rosenquist, & Kern, n.d.)

District	Grade	Number of Standards (pre-CCR)	Number of Standards (CCR)	% Change
B	6	38	59	55.3%
B	7	43	50	16.3%
B	8	43	52	20.9%
D	6	37	29	-21.6%
D	7	38	24	-36.8%
D	8	39	28	-28.2%

Figure 11. Number of math standards, by grade, by district; pre-CCR standards compared to CCR standards.



It is plausible that, in District B, the increase in the number of learning standards may have helped to weaken the relationship between ambitious teaching in the classroom and student learning gains on the state tests. Other research has documented that ambitious goals for teaching and learning mathematics are often stymied by the pressure on teachers to cover an extensive array of learning standards to prepare students to perform well on standardized tests (Banilower, Boyd, Pasley, & Weiss, 2006; Battista, 1999; Boaler, 2002; Gutstein, 2003; Horn, 2006). Math lessons associated with more ambitious learning goals – which frequently feature scenario-based problems with non-obvious and/or multiple solution strategies, small-group work, and whole class discussion – often take more class time to complete compared to more traditional direct instruction approaches (Battista, 1999; Keiser & Lambdin, 1996; Leong & Chick, 2011; Manouchehri, 1998). Consistent enactment of conceptual math lessons, with their greater demands on time, can therefore result in coverage of fewer standards over the course of a

school year, which may in some cases be reflected as depressed scores on standardized tests (Battista, 1999; Boaler, 2002). It remains a plausible hypotheses worthy of investigation in future research: even if assessments change the extent to which they measure higher-order thinking, if additional standards are added to both the intended and tested curriculum, the relationship between rigorous, conceptual instruction in the classroom and student performance on these assessments might be weakened, given the time requirements of teaching and learning mathematics in a deeper and more conceptual way. The tradeoffs between broader familiarity with numerous mathematical topics and deeper understanding of fewer key mathematical concepts seems to have been recognized and addressed by architects of the college- and career ready standards in mathematics (Conley, 2014; Porter, McMaken, Hwang, & Yang, 2011). However, it might not have been regarded as an important part of the implementation of a CCR standards-based education strategy by the policymakers in District B's state.

Limitations

The limitations to the analysis discussed here fall into two categories: (1) The inability to control for time-varying confounding factors and (2) the lack of precision in these estimates.

Longitudinal Data and Time-Varying Confounding Factors: Limitations and Advantages

The longitudinal nature of these data presents some constraints in answering the research question, which seeks to quantify the extent to which college- and career ready accountability assessments are more or less sensitive to ambitious instruction of middle school mathematics. Other notable studies which have investigated similar research questions have a markedly different data structure, utilizing scores from different assessments which were administered to the same students assigned to the same teachers over a given period of time. These analyses examined (a) how value-added estimates derived from different student assessments varied across and within teachers in a given year (e.g. Lockwood et al., 2007; Papay 2011), or (b) how the relationship between classroom observation measures and value-added changed depending on which assessment was used to estimate teacher value-added (e.g., Ing, 2017; Kane & Staiger, 2012; Grossman, Cohen, Ronfeldt, & Brown, 2014).

One advantage of these research designs – which use data from different assessments administered in a single year – is that there is less potential for unobserved time-varying confounding factors to bias the estimated relationship between observed instructional quality and teacher value-added derived from student test scores. Various analyses have suggested that a number of time-varying factors could confound this relationship. The time-varying factors which are mentioned in the literature but which are unmeasured in the current analysis include year-to-year changes in class size (Angrist & Lavy 1999; Sass, 2008), teacher-school match (Jackson, 2013), teacher peer-effects (Jackson & Bruegmann 2009), changes in grade-level content standards (see Papay 2011), teacher-student contact time, changes in the mechanism of peer effects, changes in ability tracking practices, and student-level progress monitoring and intervention practices (Sass, 2008). From our research partnership with these districts, we know some of these kinds of factors were at play in these schools. For example:

- In some of these districts and schools, struggling students were assigned to an additional mathematics course to provide tutoring services and additional "time on task" for mathematics learning. Not only did the model, features, and impact on end-of-the-year test scores vary from school to school, but the ways which these programs were implemented also changed over time (Schmidt, 2013).
- In both districts, changes in accountability assessments were accompanied by changes in the math content standards for each grade, which might have moderated the influence of teacher experience. For example, a teacher who has had several years of experience teaching the seventh grade math content will not be able to leverage that experience in the same way when the seventh grade math content changes.
- Beginning in Year 5, District D shifted to a policy of decentralized, school-based curricular adoption (Appelgate & Rosenquist, 2016).
- The nature and severity of the accountability consequences attached to state test results varied over time. In District B, state and district leaders communicated in Year 4 that results from the first year of the new assessment (Year 5) were not going to be used for school-level accountability, in order to give teachers and school leaders time to adjust to new standards and assessments. As a

consequence, assessment results from Year 4 were said to be even more important for accountability, in that schools' accountability rating would be based on Year 4 test results and that these ratings would be retained for two years (Sampson, 2010).

There may be some evidence that these or other time-varying factors could have influenced the relationship between classroom measures and value-added measures. When we compare correlation coefficients from our data with those reported from the first year of the MET Project, we see that some coefficients from our data vary as much or more within district year-to-year than the MET correlation coefficients do within a year between districts (see Figure 5 and Appendix D). Especially volatile are the relationships in student-level by-year regressions in District B, where teaching at the 85th percentile on the IQA composite measure is associated with student test score gains of 48 additional days of instruction[21] in Year 3, but where students of teachers with those same IQA scores were estimated to **underperform** similar students by equivalent amounts in Years 4 , with the difference in these coefficients in adjacent years statistically significant at the 0.01 level (see Appendix D).

While the longitudinal nature of these data do expose the estimates to bias given the potential for confounding and unobserved[22] time-varying factors as described above, the longitudinal nature of these data and analysis also confers some benefits. One reason to prefer the research design of this current dissertation analysis for investigating differences in value-added between tests is that, in both districts, the tests were state-mandated assessments with accountability consequences attached to them. In some of the analyses referred to previously, researchers examined differences in value-added (or the relationship between value-added and classroom observation measures) between state-accountability tests and lower-stakes supplemental assessments (e.g. Kane & Staiger, 2012; Papay, 2011; Sass, 2008) or between different sections of a low-stakes assessment (Papay, 2011). Research suggest that the lower stakes attached to some assessments may be associated with lower student motivation, which

[21] As in prior sections, I use a conversion rate of 0.31 standard deviations per 180 days of instruction, based on data specifically from middle-school mathematics assessments, as described in Hill, Bloom, Black, & Lipsey, 2008

[22] While some of these factors were "observed" in these sense that they were noted by researchers in the course of the research project, they are here described as contributing to "unobserved heterogeneity." In the econometric sense, "unobserved heterogeneity" is the term used for nonrandom factors between units of analysis that could add predictive power to the model but are not included because they are not adequately or systematically measured (see Zohoori and Savitz (1997) for an explanation).

manifests as behavior such as reduced persistence (less time spent on items or the test overall) and increased guessing, both of which can distort the psychometric properties of test scores and impact the validity of inferences based on these scores (Finn, 2015). In their analysis of MET Project data, Kane and Staiger (2012) affirm that the accountability context of assessment is important to consider when interpreting scores, suggesting specifically that value-added from lower-stakes assessments may not be directly comparable with value-added from state accountability assessments, stating that "value-added estimates based on state tests are in part driven by aspects of teacher performance that are specific to the state test and that may not generalize to other student outcomes of interest" (p. 12).

Another important advantage of the research design of the current study is that, while longitudinal analysis may be influenced by unobserved time-varying confounding factors, longitudinal data do allow for statistical approaches which focus on within-teacher variation in outcomes over time. Specifically, longitudinal data allows for fixed-effects analysis which remove from the analysis the time-invariant differences between teachers and focus on explaining within-teacher variation over time. In these models, we can make inferences about the connection between changes in teachers' practice and outcomes over time and provide more evidence for causation: in the fixed-effects models presented here, we can interpret results as describing how changes in a teacher's practices tended to coincide with changes in his or her value-added estimates. Because they are more suggestive of causation, these kinds of analyses produce results which are particularly relevant for informing practice; when we find that teachers who change their classroom instruction in a certain way tend to have students who outperform similar students on the state tests, we can recommend these kinds of changes to instruction to other teachers who are seeking to improve their students' performance on these tests. However, while results from teacher fixed-effects analysis may have added import in terms of relevance to practice and policy, results from the fixed-effect analyses of these data here tended to estimate relationships between IQA and value-added which were generally weaker (i.e., closer to zero) and less precisely estimated (i.e., with larger standard errors). For this reason, while results from teacher fixed-effects analyses are potentially more interesting in general, results from the teacher fixed-analyses presented here were less interesting and relatively inconclusive, especially

compared to the statistically significant results from pooled-OLS analysis in District D, which seeks to explain both within- and between-teacher effectiveness over time.

While analyses which exclusively seek to explain variations of within teacher effectiveness over time may provide stronger evidence of causality, analyses which seeks to explain both within- and between-teacher effectiveness over time are still both important and policy-relevant. The research literature suggests that a substantial portion of teacher effectiveness – anywhere between 20 to 70 percent –is between teachers and relatively stable over time (Arronson, Barrow, & Sanders, 2007; Ballou, 2005; Goldhaber & Hansen, 2013; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009) and often underestimated due to measurement error (Goldhaber & Hansen, 2013). To ignore this sizeable source of variation would limit our understanding of teacher effects as a whole. For this reason, the research community has produced (see Wayne & Youngs, 2003) and continues to produce (e.g. Harris & Sass, 2011) research which seeks to identify time-invariant characteristics of teacher effectiveness, with the rationale that this kind of research can inform policy, including both strategic human capital decisions like recruitment and retention (Goldhaber & Hansen, 2013) as well as teacher training and qualification policies (Harris & Sass, 2011).

Lack of Precision in Estimation

Lack of precision in these estimates can be attributed principally to three sources: (1) measurement error in quantifying the quality of classroom instruction, (2) a relatively small sample size of teachers, and (3) measurement error in student achievement scores.

Measurement error in quantifying the quality of classroom instruction. The classroom instruction rubric scores used here are taken from only two days of observation from a single class section. Construction of an IQA composite measure utilized factor analysis of the entire data set of 434 teacher-year observations with two sets of IQA rubric scores for each teacher-year observation. This statistical process attempts to partition error variance in the rubric scores from the shared variance across rubric scale scores and days of observation to estimate a score which only reflects the shared variance between these measures and substantially reduces the contribution of measurement error to the estimate (Costello & Osborne, 2005). In general, presence of measurement error which is independent of both the dependent and independent

variables will attenuate bias the estimates of the relationship towards zero (Wooldridge, 2005), as is confirmed in the analysis of replicated sets of simulated data described in Appendix K. In this sense, while measurement error may have contributed to type II error resulting in the lack of the rejection of the null hypothesis in District B data, conversely, it may also be that this sort of measurement error resulted in the underestimation of the statistically significant relationships identified in the District D data from the CCR years.

At the same time, even with statistical procedures which seek to minimize idiosyncratic error variation in measurement, data from two observations may not be sufficiently representative of a year's worth of instruction received by the student. A teacher may teach as many as 5 to 7 sections a day for 180 days a year. In many cases, teachers may teach more than one course (e.g., Grade 7 Math, Grade 8 Math) and the character of instruction may vary within teacher between courses. Other research suggest that the kind of rigorous mathematics-specific instructional practice measured here have greater within-teacher, day-to-day variation than other, more content general instruction practices, such as classroom management and organization (Praetorius, Lenske, & Kelmke, 2012). Also (as described in the data section), these measures were taken with some degree of teacher forewarning, and all in the spring. There may be systematic differences between teachers as to how they prepare and enact instruction when they dictate the schedule for observation. It is plausible that teachers did not select a section at random for observation but, instead, were more likely to choose classes where students were more compliant, motivated, and/or learning more quickly or more deeply than students in other sections they taught.[23] Time of year may also play a part in determining the generalizability of the observed classroom instruction: some teachers' instruction in the spring, months before end-

[23] Appendix E details results from analyses in which student-level data used for teacher value-added estimates are restricted only to students in classes which were observed and scored by MIST researchers, as opposed to utilizing data from all students associated with participating teachers in a given year, as in the primary analysis presented here. There were interesting differences compared to results from the primary analysis. Briefly, in the pre-CCR years, IQA was estimated to be much more strongly associated with student test score gains in both District B and D. After the transition to CCR standards, the relationship became strongly negative in District B, with small and inconsistent changes to the relationship in District D. However, given that teachers were able to choose the class section that was observed, and that there are very likely systematic differences between student in these classes and the classes not selected for observation, the generalizability of these results would be limited to the assessment outcomes of students in classes teachers select for observation, which is of limited policy and research interest when compared to research which examines the learning outcomes of the students of all of a teacher's classes.

or-year testing, may be more or less representative of classroom instruction across the school year as a whole.

Relying on the IQA measures alone to describe the kind of classroom instruction likely to influence student tests score gains may have also introduced some measurement error. The IQA instrument focuses on instruction in mathematics, and in particular an approach to cultivating skills of problem-solving and justification in a *launch-explore-discussion* lesson format (Boston, 2012). However, educational researchers disagree on the extent to which both content-general or content-specific classroom teaching and learning behaviors are necessary to attend to when describing quality of instruction and predicting changes in students' learning of mathematics (Schlesinger & Jentsch, 2016). Theory and empirical evidence suggest classroom management and a supportive classroom environment are two content-general aspects of classroom environments which predict student learning in mathematics and motivation for learning mathematics (Ing, 2017; Lipowski et al., 2009; Polikoff, 2014). For example, Polikoff (2014) found that the Framework for Teaching's subscale for teacher management of student behavior had the highest correlation with teacher value-added among all the subscales of all the classroom observation tools used in the MET study. A number of studies which attempt to quantify both mathematical rigor in the classroom as well as evidence of classroom management and organization have found that these two dimensions of classroom instruction are positively correlated in practice (Booker, 2014; Matsumura, Slater, & Crosson, 2008). Given these relationships, failing to collect and include measures of classroom management and organization in this analysis may have introduced bias in our estimates of the relationship between rigorous conceptual teaching and learning of mathematics in the classroom and year-over-year student test score growth.

A Relatively small sample of teachers. The per year-district sample size of teachers used here was relatively small, limiting the likelihood of generating consistent and statistically significant results. In the final analytical sample, the per district-year sample size averaged 25.4 teachers in the pre-CCR years, and 38.5 teachers in the CCR years. By contrast, the average sample size per district-year in the MET Project data utilized in Polikoff (2014) was 155.4 math teachers. Given the determinants of statistical power (Wooldridge, 2005), the negative impact of a small teacher sample size is compounded by the potential for measurement error in the teacher-

level independent variable of interest, the IQA measure of teaching quality derived from classroom observations. For this analysis, a larger sample size would have resulted in more statistical power and may have resulted in more consistent year-over-year estimates and/or additional statistically significant estimates.

Measurement error in student achievement scores. Some theoretical and empirical work has called attention to the potential for test measurement error to bias estimates of teacher value-added (Boyd, Lankford, Loeb, & Wyckoff, 2013; Herrmann, Walsh, & Isenberg, 2016; Koedel, Leatherman, & Parson. 2012; Lockwood & McCaffrey, 2014). Much of the potential for test measurement error to introduce bias in teacher value-added estimates comes from scenarios where students are sorted on true prior achievement (Lockwood & McCaffrey, 2014). This is a problem, in part, because measurement error is larger for students on the lower or higher ends of achievement for a given test (Herrmann, Walsh, & Isenberg, 2016). However, additional literature suggests that the bias introduced into teacher value-added estimates by test measurement error is relatively small and can be lessened by the introduction of statistical controls. For example, Lockford and McCaffrey (2014) explained that models similar to those employed in this dissertation, which control for aggregate prior achievement at the class- or teacher- level, can to some degree correct for test measurement error in the prior score at the individual student level. Addressing the topic more generally, Koedel, Leatherman, and Parson (2012) cited a number of analyses to provide evidence for what they characterize as a growing research consensus that the bias in teacher value-added estimates are generally small. For that reason, Koedel and colleagues and instead focus on improving the efficiency of estimates that can be achieved by accounting for test measurement error, but find these improvements to be relatively modest. Similarly, Herrmann, Walsh, and Isenberg (2016) employed shrinkage estimators to adjust the scores of students who taught large proportions of students who scored very high or very low on the achievement tests, but found that these adjustments did not produce significant differences in the likelihood that these teachers would be classified differently by the accountability system as a result of these statistical adjustments. Furthermore, while much of the theoretical and empirical work bringing attention to test measurement error has been motivated by the possibility that test measurement error will result in misclassifying teachers operating within teacher accountability systems (Herrmann, Walsh, & Isenberg, 2016; Kodel,

Leatherman, and Parsons, 2012), these concerns are less pertinent in this dissertation. While it may be more important to attend to measurement error in research which investigates the relationship between teacher effectiveness estimates and teacher accountability consequences, it may be less important to attend to this potential source of bias in research looking to explain broader trends in teaching and learning (Harris, 2009), as in this dissertation.

Implications for Policy and Practice

Policy relevance

The review of literature included in this dissertation manuscript describes some of the ways in which tests influence instruction while also making an argument that assessments which are aligned to the educational system's explicit goals for teaching and learning are necessary to realize the benefits of standards-based education reform. Results from this analysis may inform policy in that they give an indication of the extent to which these tests are aligned with ambitious goals for teaching and learning math and can therefore be validly used to measure and reward ambitious teaching and learning. In an analysis with similar research questions, Polikoff (2014) found the relationship between quality of instruction (as measured by classroom observation) and teachers' estimated value-added to be so weak as to cast doubt on the likelihood that value-added estimates from these kinds of assessments can be used to inform instruction and instructional decisions in any meaningful way. It may be that, if these student assessments cannot be used both for measuring student understanding and for deriving valid measures of ambitious teaching, then their use needs to be limited to purposes for which these test results are in fact valid. Furthermore, given the logic of standards-based education reform and the influence of assessments on instruction more generally, policymakers who would hope to promote ambitious goals for teaching and learning mathematics may need to advocate more strongly for student assessments which effectively measure and promote this kind of teaching and learning.

Policymakers and administrators judge the quality of instruction using both classroom observational measures as well as student test score gains (Goldring et al., 2015; Cohen-Vogel, 2011). We expect these measures to correlate with and reinforce each other, and the theory of change of standards-based reform is predicated in part on classroom instruction being aligned

with the kind of student learning measured in the accountability assessments. When these two kinds of measures are positively but only very weakly correlated, it presents a conundrum for those who would seek to formulate recommendation regarding how these tools should be used to improve the K-12 education system. Faced with this dilemma, researchers have responded with differing policy recommendations.

Strategy 1: Change the tests so that they align better, empirically, with measures of quality classroom instruction. The low correlations between these two kinds of measures can be interpreted as additional justification of the body of theory and empirical evidence which suggests that (1) test designed to measure student knowledge may be very different from those which would more ideally and more precisely measure quality of teaching (or teachers' contributions to student learning) such that, as a consequence, (2) most assessments used for accountability purposes are justifiably characterized as "insensitive to instruction."
The expectations for what tests can be and do is high. In 2013, Linda Darling-Hammond and 19 other educational theorists and researchers authored a document entitled *Criteria for High Quality Assessment,* which proffered a number of design principles for student assessment that might complement the adoption of new career- and college ready standards. Included among these criteria were a focus on higher-order cognitive skills and assessments' sensitivity to what happens in the classroom (i.e. that these test be "instructionally sensitive"). Test developers have made much progress on the first criterion. Darling-Hammond and colleagues (2013) cited evidence which suggest that, in the Common Core linked math assessments developed by that PARCC and Smarter Balanced , the percentage of items that required higher-order skills of analysis, synthesis, or problem solving was 70 percent, compared to 7 percent in a sample of 17 pre-CCR state math assessments.

In contrast, much less progress has been made in designing tests which are more instructionally sensitive, despite research drawing attention to this issue (e.g., D'Agostino, Welsh, & Corson, 2007; Darling-Hammond et al., 2013; Polikoff 2010, 2014; Popham 1999, 2007; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). In general, these calls for different or improved assessments may be considered a subset of a broader policy recommendation: that more resources should be allocated to improve test design and implementation given that assessments are a key driver in the systemic reform model but represent a relatively small share

of current educational spending in dollar terms (Alliance for Excellent Education, 2014; Chingos, 2013, 2015; Topol, Olson, & Roeber, 2012).

Strategy 2: Keep the tests, but de-emphasize teacher observations: This view characterizes classroom observations as expensive, time-consuming, not predictive of student learning, potentially biased, and generally producing results which differentiate very little between teachers and therefore not particularly useful and providing insufficient return on given their financial cost (Dynarski, 2016).

Strategy 3: Utilize multiple measures – This view holds that both test-score derived value-added and classroom observational measure provide complementary information and can be combined to form a more holistic and stable measure of teacher quality (Kane & Staiger, 2012).

Abandoning one of these measures in favor of the other – as suggested by Strategy 1 or 2 – would represent a significant departure from the theory of action of standards based reform which is predicated on a coherent and aligned system of measurement, supports, and accountability (O'Day & Smith, 1993). Furthermore, accountability systems which place too much value on a single measure are prone to "gaming" processes which distort the usefulness of the measure (Campbell, 2010). In contrast, the use of multiple indicators of educational quality is less likely to incentivize unintended behavior and also improves the validity of inferences about educational quality made from the data (Linn, 2000). Given the variation over time in alignment for this study's measures, one takeaway for practice may be the need for educational systems to monitor and improve the alignment and validity of their measures of quality in an ongoing way (Hill, Kapitual, & Umland, 2011). This activity could be incorporated in a larger effort to monitor the systemic impact of these measures, which would include vigilance around unanticipated undesirable responses, including "gaming" behaviors (Fredrickson & Collins, 1989; Linn, 2000; Messick, 1989).

Along with this argument for utilizing multiple *types* of measures for accountability purposes, the data and analysis here suggest a case for utilizing *repeated* measures from any given quality indicator. This would be especially important in the case of noisy measures or those which lack stability over time. Given the year-to-year instability we see in value-added estimates when we attempt to cross-validate them against our measures of teaching quality in

these data, we might argue that value-added approaches which average teacher value-added estimates over more than one year (Greene, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009) or which otherwise account for year-to-year "drift" (Chetty, Friedman, & Rockoff, 2011) would be most appropriate for policy and accountability purposes.

Implications for Research

This analysis provides some evidence that the college and career ready state accountability assessments used for middle school mathematics in District D were sensitive to ambitious math instruction, where previous versions of assessment aligned to different standards were not. This case and others like it may merit greater research, given that accountability assessments which are sensitive to content-specific measures of ambitious instruction are relatively rare in practice (Ing, 2017; Polikoff, 2014), along with the fact that the conceptual and empirical development of instructionally sensitive assessments is still ongoing (Ruiz-Primo et al., 2012).

Aside from these findings which address the initial research question, the data and analysis here also reveal other patterns worthy of further investigation. The analysis which addresses my research question draws on coefficients which essentially average the estimated relationship between classroom observation scores and teacher value-added measures across different sets of years. However, when we look at these relationships year over year, we see a great deal of fluctuation, most markedly in District B. While relatively tangential to the research question, the variation of these relationships within-district from year to year presents one of the more potentially interesting findings from these analyses.

Other studies using similar methodology but utilizing data from a single year of instruction have noted sizable differences between districts when estimating the relationship between measures of the quality of classroom instruction and teacher value-added. Polikoff (2014) found that correlations between scores on a number of classroom observation rubrics and value-added from state assessments varied substantially between districts. However, Ing (2017) analyzed these same data and found that the patterns of differences across districts were similar even when a common assessment was utilized across these districts. Because these patterns remained the same when looking at results from a common assessment, it seems that the

difference between districts' results cannot be attributed to differences in assessments (as Polikoff (2014) concluded), but are more likely due to contextual factors. This conclusion resonates with a familiar and difficult challenge of education research more generally: that the mechanisms of the policy and practice of education are frequently context dependent, making it often exceedingly difficult to replicate results and generalize findings across settings (Berliner, 2002; National Research Council, 2002, 2004a). A National Research Council report (2004a) describes education as occurring within an interaction among institutions, communities, and families and subject to physical, social, cultural, economic, and historical contextual factors which influence results in significant ways; because teaching and learning are so complex and because the US education system is so heterogeneous, it is difficult to predict the extent to which theories and findings will generalize across these contexts. The context dependence of findings in educational research – along with the often stark differences in populations and policies in school districts across the country – is familiar enough to those versed in educational research that this inter-district variation is often not a key feature in the discussion sections of some of the analyses which find and describe this kind of variation (e.g. Ing, 2017).

While the present analyses of these data did find interesting inter-district variation in these relationships, there was as much or more variation in these relationships within district over time. Some of the potential time-varying confounding factors which might have contributed to this variation were discussed in the limitations section. A few studies have described situations in which changes in district policies, practices, and populations have resulted in change of the influence of key variables of interest over time. For example, Lemons, Fuchs, Gilbert, and Fuchs (2014) described a series of randomized controlled trials of a supplemental reading program for students in a single school district in Grades 2 to 5 and found that differences between treatment and control group across a number of student outcomes shrank and eventually became statistically insignificant, most likely due to changes in school population, policies, and practices over the course of nine years. In their study of a number of districts' efforts to improve teaching and learning in the science, technology, engineering, and math (STEM) fields, Banilower and colleagues (2006) described how a number of contextual factors – especially teacher turnover, instructional coach turnover, principal turnover, superintendent turnover, new initiatives, and competing priorities – influenced the

implementation of efforts to improve teaching and learning. While these kinds of staffing and policy issues do not in themselves constitute likely or potential unobserved time-varying factors which confound the estimated relationship between observed instructional quality and teacher-value added from student test scores, it is plausible that these staffing and policy changes have the potential to lead to the kinds of confounding factors enumerated previously in the limitations section (e.g., changes in class sizes, tracking practice, progress monitoring, and invention practices).

Social scientists have long recognized the phenomenon that change in conditions over time often leads to change in the relationship between different variables. In 1975, Cronbach wrote of how "generalizations decay" and after time explain little variation, especially in the study of complex social phenomena (pp 122-123). What is important to note here – and what some of the data and analysis here suggest – is that that relative to the amount of variation existing between different school districts, the amount of variation within school districts across time may be larger than previously conceived. Given the charge of the quantitative social scientist to understand and attempt to explain variation in processes and outcomes, the sizeable variation occurring within school districts over time may merit greater attention.

In future research which might seek to explain variation within a given context over time with greater depth and subtly, it may be important to consider the choice of value-added methodology. Just as I suggested above that approaches which use multiple years of teacher-value-added estimates to reduce intertemporal variability of teacher value-added estimates (Chetty, Friedman, & Rockoff, 2011; McCaffrey, Sass, Lockwood, & Mihaly, 2009) might be most appropriate for policy and accountability purposes, these kinds of approaches might not be appropriate for all kinds of research questions, especially those which might attempt to explore and explain interesting and important intertemporal variation.[24] In this sense, the empirical results described here lend support to a framework articulated by Harris (2009), that the assumptions and measurement characteristics needed for value-added estimates used for research and program evaluation purposes are very different from those necessary for accountability purposes.

[24] For more detail on why the approach to estimating teacher value-added employed by Chetty, Friedman, & Rockoff (2011) is not well suited to this research question and data, see Appendix G

APPENDIX A

Adoption of Common Core or "College- and Career Ready" Standards in Mathematics

According to a September 2017 update, 35 states and the District of Columbia had previously adopted and retained the Common Core State Standards in Mathematics (CCSS-M; Ujifusa, 2017; See Figure A1). Ten states have "announced a major Common Core rewrite or replacement"; four have never adopted the Common Core State Standards, and one has adopted the Common Core Standards in English Language Arts only.

For the purposes of this dissertation, I searched for documents from the respective Departments of Education or state legislatures in the fifteen states which are not presently implementing the Common Core Standards in Mathematics, looking for evidence as to whether these standards explicitly aspire to promote "College- and Career Readiness" (Table A1). Some explicit reference to the promotion of "College- and Career Readiness" – sometimes capitalized, sometimes not – was found in fourteen of these fifteen states.

Of these fifteen sets of state mathematics standards, the characterization of "College- and Career- Readiness" was least explicit in Minnesota, where the standards were described as being influenced by "College and *Work* Readiness Expectations" [emphasis added], authored by the Minnesota P-16 Education Partnership Working Group. At the same time, the standards from the American Diploma Project of Achieve, Inc., was cited as one of the sources of the standards. The standards from the American Diploma Project were also influential in the formulation of the CCSS (Conley, 2014), suggesting the possibility of some alignment to the CCSS-M and college- and career ready goals more broadly. However, the current Minnesota standards in mathematics were adopted in 2007 and have not been revised since. As such, the current Minnesota state standards were adopted before the formulation and circulation of the public release of the draft of college and career ready standards by the National Governors Association and the Council of Chief State School Officers in September of 2009 (Common Core State Standards Initiative, n.d.). For this reason, a textual analysis comparing the CCSS-M and the current Minnesota – which is beyond the scope of this dissertation – may find less alignment of these two sets of standards. It may be that more recently adopted non-CCSS-M standards are more likely to be

informed by the well-circulated CCSS-M standards. This may be the case for many of the standards of the 15 states not currently implementing the Common Core Standards, given that more than half of these states adopted their standards in 2016 or later.

While the Texas standards in mathematics were also adopted prior to September 2009, I was able to identify additional evidence of both explicit promotion of College and Career Ready goals, as well as evidence of some alignment with the CCSS-M (Conley et al., 2011)

Figure A1. Status of Common Core Standards adoption, by state/federal district, as of September 2017 (source: Ujifusa, 2017).

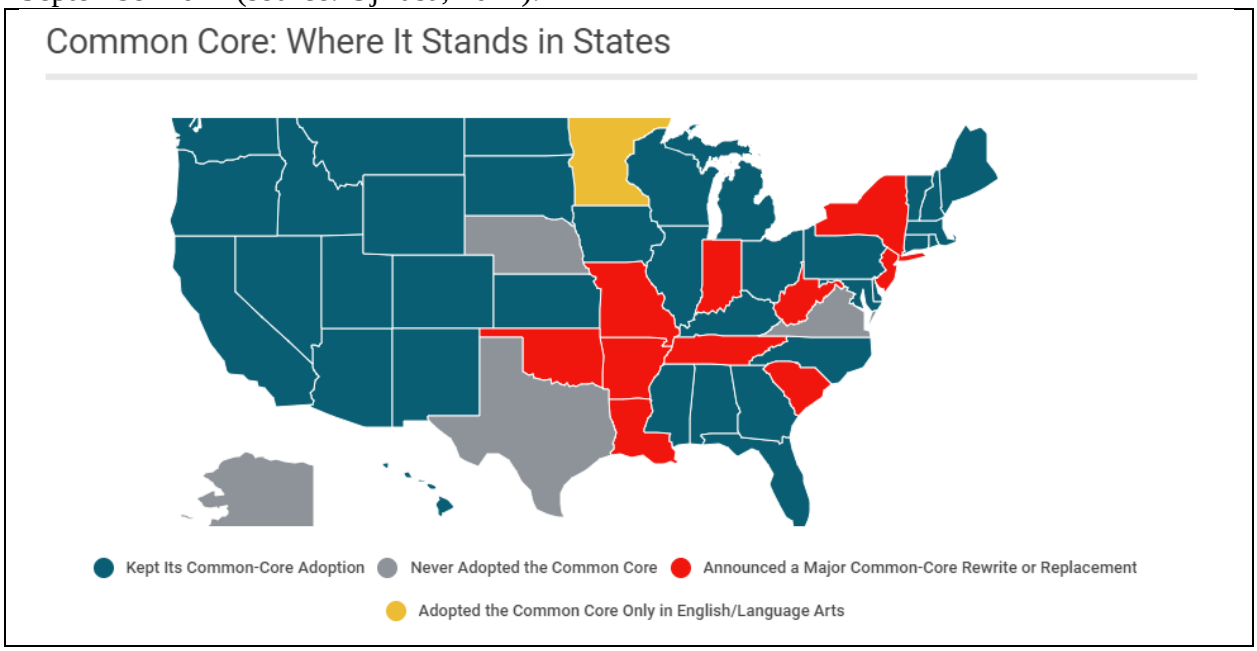


Table A1. For states not adopting and retaining the Common Core State Standards in Mathematics, evidence on date of adoption and explicit reference to "college- and career readiness".(Part I)

State/Federal District	Status	Documents explicitly reference College- and Career Ready objectives	Year of authorship or adoption	Reference (Retrieved from World Wide Web resources, November 2017)
Minnesota	Adopted the Common Core Only in English/Language Arts	Yes (qualified)	2007	http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=005246&RevisionSelectionMethod=latestReleased&Rendition=primary
Arkansas	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	http://www.arkansased.gov/public/userfiles/Learning_Services/Curriculum%20and%20Instruction/Frameworks/Math/Arkansas_Mathematics_Standards_K_5.pdf
Indiana	Announce a Major Common-Core Rewrite or Replacement	Yes	2014	https://www.doe.in.gov/sites/default/files/standards/mathematics/grade-6-standards.pdf
Louisiana	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	http://www.louisianabelieves.com/docs/default-source/academic-standards/louisiana-state-standards-(ela-math).pdf?sfvrsn=10
Missouri	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	https://dese.mo.gov/college-career-readiness/curriculum/missouri-learning-standards
New Jersey	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	http://www.state.nj.us/education/cccs/2016/math/standards.pdf
New York	Announce a Major Common-Core Rewrite or Replacement	Yes	2017	http://www.nysed.gov/common/nysed/files/ela-and-mathematics-standards-preface.pdf
Oklahoma	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	http://sde.ok.gov/sde/sites/ok.gov.sde/files/OAS-Math-Final%20Version_3.pdf
South Carolina	Announce a Major Common-Core Rewrite or Replacement	Yes	2015	https://ed.sc.gov/instruction/standards-learning/mathematics/standards/scccr-standards-for-mathematics-final-print-on-one-side/

Table A2. For states not adopting and retaining the Common Core State Standards in Mathematics, evidence on date of adoption and explicit reference to "college- and career readiness". (Part II)

State/Federal District	Status	Documents explicitly reference College- and Career Ready objectives	Year of authorship or adoption	Reference (Retrieved from World Wide Web resources, November 2017)
Tennessee	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	https://www.tn.gov/assets/entities/sbe/attachments/4-15-16_V_A_Math_Standards_Attachment.pdf
West Virginia	Announce a Major Common-Core Rewrite or Replacement	Yes	2016	http://apps.sos.wv.gov/adlaw/csr/readfile.aspx?DocId=27353&Format=PDF
Alaska	Never Adopted the Common Core	Yes	2012	https://education.alaska.gov/akstandards/standards/akstandards_elaandmath_080812.pdf
Nebraska	Never Adopted the Common Core	Yes	2015	https://www.education.ne.gov/wp-content/uploads/2017/07/2015_Nebraska_College_and_Career_Standards_for_Mathematics_Vertical.pdf
Texas	Never Adopted the Common Core	Yes	2008	http://www.thecb.state.tx.us/collegereadiness/CRS.pdf
Virginia	Never Adopted the Common Core	Yes	2011	http://www.doe.virginia.gov/instruction/mathematics/capstone_course/perf_expectations_math.pdf

APPENDIX B

IQA Rubrics

Table B1. IQA Rubric 1: Potential of the Task: Did the Task Have Potential to Engage Students in Rigorous Thinking about Challenging Content? (source: Boston, 2012)

4	**The task has the potential to engage students in exploring and understanding the nature of mathematical concepts, procedures, and/or relationships, such as:** • Doing mathematics: using complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); OR • Procedures with connections: applying a broad general procedure that remains closely connected to mathematical concepts. The task must explicitly prompt for evidence of students' reasoning and understanding. For example, the task **MAY** require students to: • solve a genuine, challenging problem for which students' reasoning is evident in their work on the task; • develop an explanation for why formulas or procedures work; • identify patterns; form and justify generalizations based on these patterns; • make conjectures and support conclusions with mathematical evidence; • make explicit connections between representations, strategies, or mathematical concepts and procedures. follow a prescribed procedure in order to explain/illustrate a mathematical concept, process, or relationship.
3	**The task has the potential to engage students in complex thinking or in creating meaning for mathematical concepts, procedures, and/or relationships. However, the task does not warrant a "4" because:** • the task does not explicitly prompt for evidence of students' reasoning and understanding. • students may be asked to engage in doing mathematics or procedures with connections, but the underlying mathematics in the task is not appropriate for the specific group of students (i.e., too easy or too hard to promote engagement with high-level cognitive demands); • students may need to identify patterns but are not pressed to form or justify generalizations; • students may be asked to use multiple strategies or representations but the task does not explicitly prompt students to develop connections between them; students may be asked to make conjectures but are not asked to provide mathematical evidence or explanations to support conclusions
2	**The potential of the task is limited to engaging students in using a procedure that is either specifically called for or its use is evident based on prior instruction, experience, or placement of the task.** There is little ambiguity about what needs to be done and how to do it. **The task does not require students to make connections to the concepts or meaning underlying the procedure being used.** Focus of the task appears to be on producing correct answers rather than developing mathematical understanding (e.g., applying a specific problem solving strategy, practicing a computational algorithm). OR **There is evidence that the mathematical content of the task is at least 2 grade-levels below the grade of the students in the class.**
1	**The potential of the task is limited to engaging students in memorizing or reproducing facts, rules, formulae, or definitions. The task does not require students to make connections to the concepts or meaning that underlie the facts, rules, formulae, or definitions being memorized or reproduced.**
0	**The task requires no mathematical activity.**

Table B2. IQA Rubric 2: Implementation of the Task: At what level did the teacher guide students to engage with the task in implementation? (source: Boston, 2012)

4	**Students engaged in exploring and understanding the nature of mathematical concepts, procedures, and/or relationships, such as:** • Doing mathematics: using complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); OR • Procedures with connections: applying a broad general procedure that remains closely connected to mathematical concepts. There is explicit evidence of students' reasoning and understanding. For example, students may have: • solved a genuine, challenging problem for which students' reasoning is evident in their work on the task; • developed an explanation for why formulas or procedures work; • identified patterns, formed and justified generalizations based on these patterns; • made conjectures and supported conclusions with mathematical evidence; • made explicit connections between representations, strategies, or mathematical concepts and procedures. followed a prescribed procedure in order to explain/illustrate a mathematical concept, process, or relationship.
3	**Students engaged in complex thinking or in creating meaning for mathematical concepts, procedures, and/or relationships. However, the implementation does not warrant a "4" because:** • there is no explicit evidence of students' reasoning and understanding. • students engaged in doing mathematics or procedures with connections, but the underlying mathematics in the task was not appropriate for the specific group of students (i.e., too easy or too hard to sustain engagement with high-level cognitive demands); • students identified patterns but did not form or justify generalizations; • students used multiple strategies or representations but connections between different strategies/representations were not explicitly evident; students made conjectures but did not provide mathematical evidence or explanations to support conclusions
2	**Students engaged in using a procedure that was either specifically called for or its use was evident based on prior instruction, experience, or placement of the task.** There was little ambiguity about what needed to be done and how to do it. S**tudents did not connections to the concepts or meaning underlying the procedure being used.** Focus of the implementation appears to be on producing correct answers rather than developing mathematical understanding (e.g., applying a specific problem solving strategy, practicing a computational algorithm). **OR** **There is evidence that the mathematical content of the task is at least 2 grade-levels below the grade of the students in the class.**
1	**Students engage in memorizing or reproducing facts, rules, formulae, or definitions. Students do not make connections to the concepts or meaning that underlie the facts, rules, formulae, or definitions being memorized or reproduced.**
0	**Students did not engage in mathematical activity.**
N/A	**The students did not engage with a mathematical task.**

Table B3. IQA Rubric 3: Student Discussion Following Task: To what extent did students show their work and explain their thinking about the important mathematical content? (source: Boston, 2012)

4	Students show/describe written work for solving a task and/or engage in a discussion of the important mathematical ideas in the task. During the discussion, students: provide complete and thorough explanations of why their strategy, idea, or procedure is valid; students explain why their strategy works and/or is appropriate for the problem; students make connections to the underlying mathematical ideas (e.g., "I divided because we needed equal groups"). OR Students show/discuss more than one strategy or representation for solving the task, provide explanations of why/how the different strategies/representations were used to solve the task, *and/or make connections between strategies or representations. [Thorough presentation and discussion across strategies or representation]*
3	Students show/describe written work for solving a task and/or engage in a discussion of the important mathematical ideas in the task. During the discussion, students provide explanations of why their strategy, idea, or procedure is valid and/or students begin to make connections BUT the explanations and connections are not complete and thorough (e.g., student responses often require extended press from the teacher, are incomplete, lack precision, or fall short making explicit connections). OR Students show/discuss more than one strategy or representation for solving the task, and provide explanations of why/how the individual strategies/representations were used to solve the task *but do not make connections between different strategies or representations. [Thorough presentation and/or discussion of individual strategies or representations (no cross-talk)]*
2	Students show/describe written work for solving the task (e.g., the steps for a multiplication problem, finding an average, or solving an equation; what they did first, second, etc) but do not engage in a discussion of why their strategies, procedures, or mathematical ideas work*; do not make connection to mathematical concepts. [Procedural explanations only]* OR Students show/discuss only one strategy or representation for solving the task.
1	Students provide brief or one-word answers (e.g., fill in blanks); OR **Student's responses are non-mathematical.**
0	There was no discussion of the task.
N/A	No class discussion

APPENDIX C

Descriptions of Main Model and Robustness Models for Estimating Teacher Value-added

Table C1 Description of specification for main value-added model (following MET Project Methodology) and 14 variations on this approach, to utilize as robustness checks.

Model	Estimation Steps	Shrinkage	Prior Achievement (y-1)	Prior Achievement (y-2)	Student-level demographics	Class-level demographics	Class-level prior achievement (y-1)	School-level demographics	School-fixed effects
Main	*2*	*fixed*	*X*		*X*	*X*	*X*		
Robustness1	2	fixed	X						
Robustness2	2	fixed	X		X				
Robustness3	2	fixed	X	X	X	X	X		
Robustness4	2	fixed	X	X	X	X	X	X	
Robustness5	2	fixed	X	X	X	X	X		X
Robustness6	2	random	X		X	X	X		
Robustness7	2	random	X						
Robustness8	2	random	X	X	X	X	X		X
Robustness9	1	fixed	X		X	X	X		
Robustness10	1	fixed	X						
Robustness11	1	fixed	X	X	X	X	X		X
Robustness12	1	random	X		X	X	X		
Robustness13	1	random	X						
Robustness14	1	random	X	X	X	X	X		X

Note: One or more of these specifications have been described in at least one of the following analyses: Ballou, Mokher, & Cavalluzzo, 2012; Corcoran & Jennings, 2012; Goldhaber, Walch, & Gabele, 2014; Lipscomb, Gill, Booker, & Johnson, 2010; Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007; and Papay, 2011

Table C2. District B: Spearman rank correlation of main value-added model, robustness models

	Main Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
Main Model	1.000														
Model 1	0.917	1.000													
Model 2	0.890	0.927	1.000												
Model 3	0.931	0.872	0.960	1.000											
Model 4	0.868	0.788	0.830	0.888	1.000										
Model 5	0.794	0.735	0.770	0.818	0.914	1.000									
Model 6	0.998	0.916	0.891	0.931	0.865	0.788	1.000								
Model 7	0.916	0.998	0.926	0.872	0.789	0.732	0.918	1.000							
Model 8	0.789	0.729	0.763	0.811	0.904	0.997	0.785	0.728	1.000						
Model 9	0.873	0.859	0.819	0.808	0.724	0.678	0.876	0.860	0.678	1.000					
Model 10	0.805	0.887	0.811	0.751	0.670	0.627	0.807	0.886	0.627	0.930	1.000				
Model 11	0.801	0.819	0.861	0.837	0.721	0.689	0.805	0.820	0.690	0.934	0.899	1.000			
Model 12	0.977	0.937	0.906	0.914	0.834	0.771	0.979	0.939	0.769	0.933	0.860	0.864	1.000		
Model 13	0.901	0.990	0.919	0.858	0.775	0.721	0.904	0.992	0.718	0.883	0.917	0.843	0.936	1.000	
Model 14	0.792	0.734	0.768	0.810	0.904	0.982	0.788	0.732	0.982	0.719	0.661	0.725	0.789	0.730	1.000
	Main Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14

Table C3. District D: Spearman rank correlation of main value-added model, robustness models

	Main Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
Main Model	1.000														
Model 1	0.892	1.000													
Model 2	0.905	0.945	1.000												
Model 3	0.965	0.864	0.938	1.000											
Model 4	0.873	0.756	0.823	0.912	1.000										
Model 5	0.795	0.701	0.767	0.845	0.929	1.000									
Model 6	0.997	0.889	0.902	0.962	0.872	0.789	1.000								
Model 7	0.893	0.998	0.945	0.866	0.757	0.701	0.893	1.000							
Model 8	0.794	0.697	0.763	0.842	0.926	0.994	0.792	0.699	1.000						
Model 9	0.832	0.823	0.845	0.821	0.708	0.628	0.832	0.823	0.627	1.000					
Model 10	0.814	0.953	0.881	0.781	0.675	0.621	0.811	0.950	0.613	0.845	1.000				
Model 11	0.810	0.802	0.871	0.844	0.732	0.651	0.809	0.803	0.650	0.962	0.820	1.000			
Model 12	0.977	0.906	0.932	0.953	0.836	0.758	0.980	0.911	0.760	0.861	0.831	0.845	1.000		
Model 13	0.867	0.994	0.934	0.840	0.726	0.672	0.868	0.996	0.671	0.811	0.963	0.791	0.891	1.000	
Model 14	0.792	0.707	0.776	0.843	0.919	0.977	0.787	0.710	0.979	0.634	0.621	0.661	0.771	0.680	1.000
	Main Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14

APPENDIX D

Details of Instability of Value-added and Classroom Observation Measures over Time and across Districts.

Table D1. Between district (and assessment) variation in correlation of value-added estimates with classroom observation measures, from the Measures of Effective Teaching (MET) study (*p≤0.05) (source: Polikoff, 2014).

	Overall	**Dist. 1**	**Dist. 2**	**Dist. 4**	**Dist. 5**	**Dist. 6**
FFT composite	0.18*	0.31*	0.03	0.13	0.26*	0.23*
n	805	85	173	184	180	183
CLASS composite	0.15*	0.18	0.04	0.08	0.19*	0.28*
n	804	85	173	183	180	183
MQI composite	0.03	-0.04	0.01	0.07	0.03	0.18*
N	794	84	166	183	178	166

Table D2. Year to year variation in correlation of value-added estimates with classroom observation measures, by district, from the current study (+ p≤0.10, *p≤0.05, **p≤0.01), all data collected fitting initial analysis criteria

	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**
IQA Composite (Dist B)	**-0.35+**	-0.15	**0.36***	-0.29	0.06	-0.14	-0.21
N	24	25	31	26	38	43	39
IQA Composite (Dist D)	0.08	0.24	0.14	-0.06	0.00	**0.26+**	**0.25+**
N	23	26	26	24	21	46	45

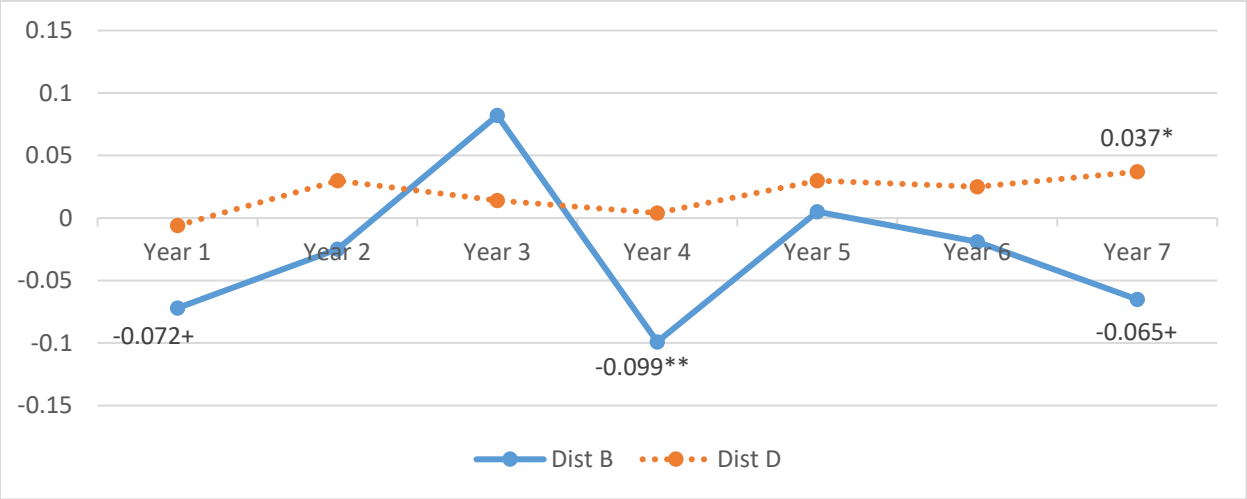
Table D3. Year to year variation in correlation of value-added estimates with classroom observation measures, by district, from the current study (+ p≤0.10, *p≤0.05, **p≤0.01), all data collected fitting final analysis criteria (3 outliers removed).

	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**
IQA Composite (Dist B)	**-0.35+**	-0.15	**0.51****	-0.09	0.06	-0.14	-0.21
N	24	25	***29***	***25***	38	43	39
IQA Composite (Dist D)	0.08	0.24	0.14	-0.06	0.23	**0.26+**	**0.25+**
N	23	26	26	24	***20***	46	45

Table D4. By year, by district IQA correlation coefficient results when IQA is included in student-level value-added specification (i.e., current year achievement predicted by classroom IQA score, controlling for prior achievement, student demographics, and class-level demographics and prior achievement). P-values calculated with clustered standard-errors at the teacher level (+ p≤0.10, *p≤0.05, **p≤0.01).

	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**
IQA Composite (Dist B)	**-0.072+**	-0.025	0.082	**-0.099****	0.005	-0.019	**-0.065+**
p-value	(0.052)	(0.449)	(0.123)	(0.005)	(0.888)	(0.323)	(0.062)
N	1749	1474	2045	1587	2576	4100	3209
IQA Composite (Dist D)	-0.006	0.030	0.014	0.004	0.030	0.025	**0.037***
p-value	(0.869)	(0.122)	(0.646)	(0.846)	(0.487)	(0.274)	(0.020)
N	2077	2430	2336	2286	1211	4491	8268

Figure D1. By year, by district IQA regression coefficient results when IQA is included in student-level value-added specification (i.e., current year achievement predicted by classroom IQA score, controlling for prior achievement, student demographics, and class-level demographics and prior achievement). Only significant or marginally significant point estimates labeled.



By pooling two adjacent years of data and interacting the IQA coefficient with an indicator variable for one of the years, I am able to conduct a statistical test to determine if the coefficient estimated for the first year is different from the following year at a level of statistical significance, or if there is a greater likelihood that the estimated differences could be ascribed to sample error. This analysis (Table D5) suggests that the estimated probabilities that the year-on-year change in the IQA coefficient in Years 3, 4, and 5 could be produced by sample error are 8.7, 0.6, and 2.8 percent, respectively. This provides a degree of compelling evidence that the change in relationship observed is unlikely to be due to sampling error but that unobserved time-varying confounding factors may play a role (see Chapter VI for a discussion). Furthermore, the estimated IQA effect sizes in District B in Years 3 and 4 approach the estimated teacher effect size in these years (see Chapter V, Figure 1), which are consistent with the prior year literature in their magnitude (e.g. Hanushek & Rivkin, 2010; Nye, Konstantopoulos, & Hedges, 2004). We would conclude then that the size of these effects are practically significant, given the literature on teacher effect sizes which is cited to describe teacher effectiveness as the most important school factor in explaining differences in student achievement (Goldhaber, 2002). This comparison – along with the evidence from the statistical tests – suggests that the year-to-year

fluctuations we see here are both statistically and practically significant, and not accurately characterized as "dancing around zero." In other words, this evidence suggests they are meaningful differences which are not likely ascribed to noise and sampling error alone.

Table D5. Difference between within-district IQA coefficients in adjacent years. P-values calculated with clustered standard-errors at the teacher level (+ p≤0.10, *p≤0.05, **p≤0.01).

	Y2-Y1 Coef.	**Y3-Y2 Coef.**	**Y4-Y3 Coef.**	**Y5-Y4 Coef.**	**Y6-Y5 Coef.**	**Y7-Y6 Coef.**
Dist B	0.048	**0.107+**	**-0.181****	**0.105***	-0.027	-0.043
(p-value)	(0.244)	(0.087)	(0.006)	(0.028)	(0.529)	(0.310)
Dist D	0.036	-0.015	-0.011	0.026	-0.004	0.012
(p-value)	(0.360)	(0.702)	(0.781)	(0.568)	(0.926)	(0.681)

As an additional robustness check, the cohort fixed effects described in Chapter III and Chapter IV were substituted with school-level effects to determine if this specification helped to explain the instability of estimates over time. On the contrary, this substitution exacerbated instability over time, generally resulting in point estimates which were generally larger and more precise, with notable increases in fluctuation in Distric D (see Table D6, Figure D2, and Table D7).

Table D6. By year, by district IQA correlation coefficient results when teacher IQA and school-level fixed effects are included in student-level value-added specification (i.e., current year achievement predicted by classroom IQA score, controlling for prior achievement, student demographics, and class-level demographics and prior achievement). P-values calculated with clustered standard-errors at the teacher level (+ p≤0.10, *p≤0.05, **p≤0.01).

	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**
IQA Composite (Dist B)	-0.062	**-0.06+**	**0.102***	**-0.088***	**-0.035***	-0.009	-0.022
p-value	(0.168)	(0.053)	(0.012)	(0.016)	(0.024)	(0.642)	(0.531)
N	1749	1474	2045	1587	2576	4100	3209
IQA Composite (Dist D)	-0.025	0.008	-0.006	0.01	**0.128*****	-0.018	**0.078****
p-value	(0.494)	(0.736)	(0.823)	(0.398)	(0.001)	(0.344)	(0.003)
N	2077	2430	2336	2286	1211	4491	8268

Figure D2. By year, by district IQA regression coefficient results when IQA and school fixed effects are included in student-level value-added specification (i.e., current year achievement predicted by classroom IQA score, controlling for prior achievement, student demographics, and class-level demographics and prior achievement). Only significant or marginally significant point estimates labeled.

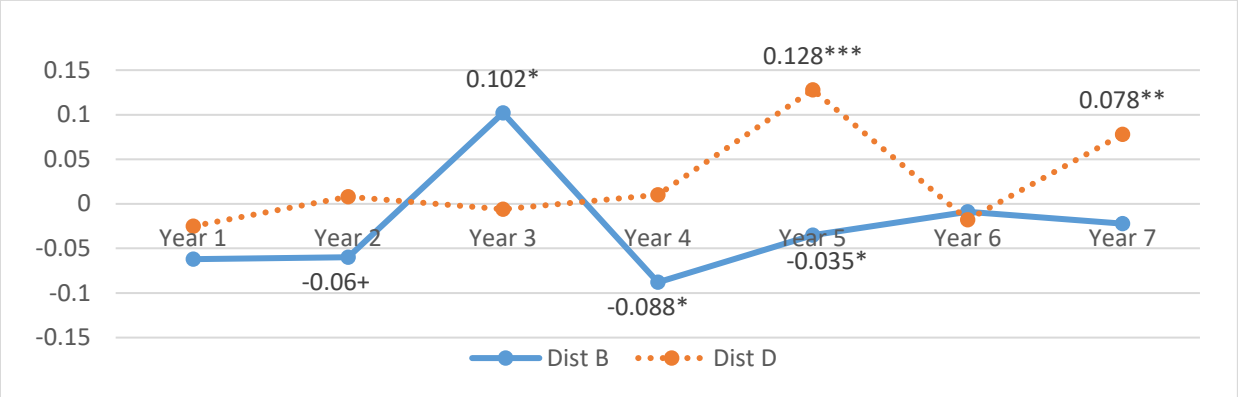


Table D7 Linear combinations of coefficients of teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years (Fully interacted, POLS regressions). (P-values in parentheses). Column 1 includes cohort fixed effects, as described in Chapter III and Chapter IV. Column 2 includes instead school-level fixed effects

	(1)	(2)
IQA Composite in pre-CCR Years (Dist B)	0.007 (0.705)	0.011 (0.641)
IQA Composite in CCR Years (Dist B)	-0.017 (0.330)	-0.017 (0.326)
IQA Composite in pre-CCR Years (Dist D)	0.011 (0.346)	0.007 (0.558)
IQA Composite in CCR Years (Dist D)	**0.048*** ** (0.000)	**0.041** ** (0.003)
Difference :IQA Coef. in Dist. D and B (Pre –CCR Years)	0.004 (0.874)	-0.003 (0.903)
Difference :IQA Coef. in Dist. D and B (CCR Years)	**0.066** ** (0.003)	**0.057** ** (0.009)
Difference: Pre-CCR to CCR Dist B	-0.025 (0.343)	-0.027 (0.334)
Difference: Pre-CCR to CCR Dist D	**0.037*** (0.032)	**0.033*** (0.048)
Difference :IQA "slopes". in Dist D and B (ΔD-ΔB)	**0.062*** (0.048)	**0.061**+ (0.066)

It may be worth noting that the one-step estimation method which produced the results reported in Table D6, Figure D2, and Table D7 are similar to those employed in Lynch, Chin, and Blazar's (2017) analysis of the relationship between the Mathematical Quality of Instruction (MQI) observation instrument and student test score gains. This analysis uses employs a single regression with school fixed effects, while Lynch and colleagues instead employ student demographic variables aggregated to the school level to control for school-level differences. However, the beta coefficients estimated in some of the districts in that analysis (0.081 and 0.126, $p<0.01$ for both) are very similar to those estimated for District D in Year 7 and Year 5, respectively.

APPENDIX E

Regression Results from Fully Interacted Models

Table E1. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. (Fully interacted mode). (P-values in parentheses).

	Districts B & D (pooled)		
	Fixed Effect	HLGC	POLS
	(1)	(2)	(3)
IQA Composite	0.026	0.015	0.007
	(0.444)	(0.372)	(0.705)
IQAxCCR	-0.050	-0.037	-0.025
	(0.211)	(0.120)	(0.343)
CCR	-0.042	-0.029	-0.019
	(0.112)	(0.415)	(0.338)
Dist. D	(NA)	0.046	0.029
	(NA)	(0.597)	(0.594)
Dist. DxIQA	-0.022	-0.002	0.004
	(0.539)	(0.935)	(0.874)
Dist. DxCCR	-0.033	-0.011	-0.013
	(0.465)	(0.837)	(0.704)
Dist D.xCCRxIQA	0.067	**0.058+**	**0.062***
	(0.135)	(0.066)	(0.048)
Year Slope (pre-CCR)		0.002	
		(0.835)	
Dist D. x Year Slope (pre-CCR)		-0.004	
		(0.827)	
Year Slope (CCR)		0.013	
		(0.418)	
Dist D. x Year Slope (CCR)		-0.021	
		(0.431)	
Cohort Fixed-Effects	X	X	X
Intercept	0.030	-0.002	0.006
	(0.255)	(0.926)	(0.734)
N	434	434	434
AIC	-1011.8	-430.9	-379.6
BIC	-983.3	-361.7	-326.6

Table E2 Linear combinations of coefficients of teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. (Fully interacted mode). (P-values in parentheses).

IQA Composite in pre-CCR Years (Dist B)	0.026 (0.444)	0.015 (0.372)	0.007 (0.705)
IQA Composite in CCR Years (Dist B)	-0.025 (0.282)	-0.022 (0.372)	-0.017 (0.330)
IQA Composite in pre-CCR Years (Dist D)	0.003 (0.814)	0.013 (0.391)	0.011 (0.346)
IQA Composite in CCR Years (Dist D)	0.020 (0.268)	**0.034*** (0.022)	**0.048***** (0.000)
Difference :IQA Coef. in Dist. D and B (Pre –CCR Years)	-0.022 (0.539)	-0.002 (0.935)	0.004 (0.874)
Difference :IQA Coef. in Dist. D and B (CCR Years)	0.045 (0.127)	**0.057*** (0.015)	**0.066**** (0.003)
Difference: Pre-CCR to CCR Dist B	-0.050 (0.211)	-0.037 (0.120)	-0.025 (0.343)
Difference: Pre-CCR to CCR Dist D	0.016 (0.397)	0.021 (0.311)	**0.037*** (0.032)
Difference :IQA "slopes". in Dist D and B (ΔD-ΔB)	0.067 (0.135)	**0.058+** (0.066)	**0.062*** (0.048)

Figure E1a-c. Comparison of Point Estimates by time period, by district, by estimation method

Figure E1a Teacher-Fixed Effects model

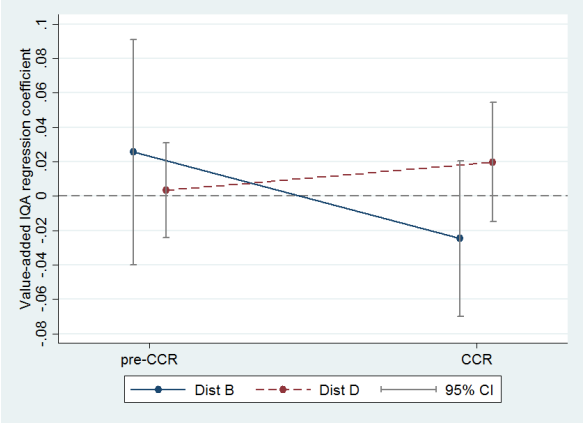


Figure E1b. Hierarchical Linear Growth Curve model

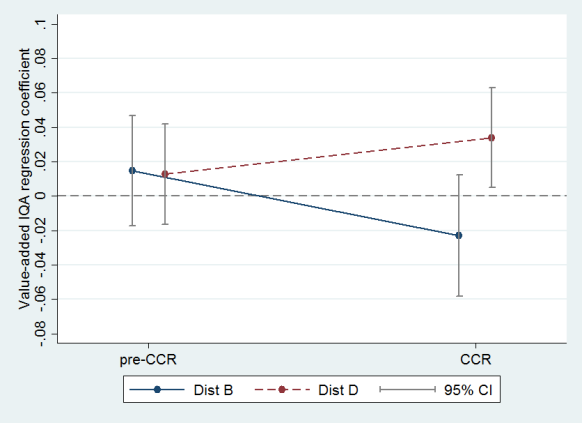
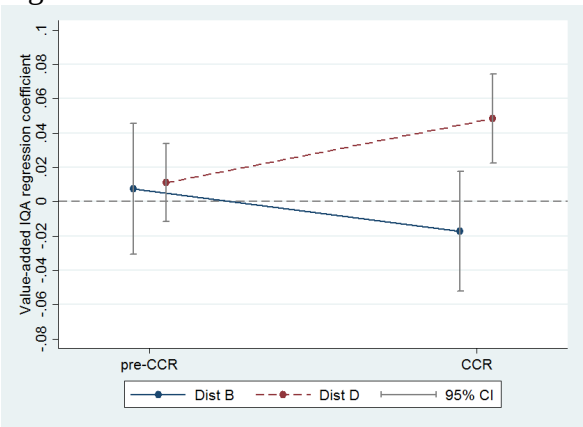


Figure E1c. Pooled-OLS model



APPENDIX F

Results from Observed Only Classes

Table F1. Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years, with value-added scores restricted only to students in classes observed by MIST researchers. (P-values in parentheses). Compare to Table 11, where value-added scores are calculated using student-level data from all students associated with a participating teacher, not only those student in observed classes. No minimum number of students. (P-values in parentheses)

	District B.			District D.		
	Fixed Effect	HLGC	POLS	Fixed Effect	HLGC	POLS
	(1)	(2)	(3)	(4)	(5)	(6)
IQA Composite (2 Day)	**0.062+**	**0.057***	**0.051***	0.030	**0.038***	**0.040***
	(0.098)	(0.016)	(0.033)	(0.152)	(0.012)	(0.011)
IQA x CCR	**-0.148*****	**-0.107****	**-0.091****	-0.034	0.005	-0.027
	(0.000)	(0.001)	(0.004)	(0.326)	(0.840)	(0.234)
CCR	0.026	0.040	0.042	**-0.150*****	-0.070	-0.032
	(0.527)	(0.421)	(0.113)	(0.000)	(0.261)	(0.346)
Year Slope (pre CCR)		0.006			-0.004	
		(0.720)			(0.794)	
Year Slope (CCR)		0.010			0.014	
		(0.638)			(0.653)	
Cohort Controls	X	X	X	X	X	X
Intercept	**-0.029***	-0.023	-0.012	-0.014	0.005	0.001
	(0.022)	(0.504)	(0.536)	(0.536)	(0.859)	(0.937)
N	218	218	218	194	194	194
AIC	-320.0	-91.2	-84.0	-402.8	-165.8	-116.6
BIC	-309.9	-67.5	-67.1	-389.7	-133.2	-90.5
IQA Composite in CCR Years (IQA coefficient + CCR interaction Term)	**-0.087****	**-0.050***	**-0.040+**	-0.004	**0.042***	0.013
	(0.007)	(0.038)	(0.072)	(0.867)	(0.020)	(0.544)

The sample size for these regressions was slightly smaller than those used for the primary estimates; The District B sample size was reduced from 224 to 218, and the District D sample size was reduced from 210 to 194. This is attributed to the fact that not all yearly observations were able to be matched to specific classes at the student-level. In order to make sure that the differences in regression results are attributable to restricting the analysis to students in observed

classes only and not to change in teacher sample, another regression was performed with using the same teacher sample in regressions from Table F1, but calculating value-added using all students associated with participating teacher, not only those in observed classes. There results were closer to those of the main models (Table 11). This suggests that the change in regression results is more attributable to change in student sample (i.e., looking at students in observed classes only) that in change in teacher sample.

Table F2 Teacher value-added estimates regressed on IQA classroom observation score and interaction term for CCR assessment years. Identical teacher sample as in Table E1, but with value-added scores calculated using all students (i.e. students in both observed and unobserved classes. No minimum number of students. (P-values in parentheses)

	District B.			District D.		
	Fixed Effect	HLGC	POLS	Fixed Effect	HLGC	POLS
	(1)	(2)	(3)	(4)	(5)	(6)
IQA Composite (2 Day)	0.025	0.015	0.008	0.004	0.016	0.017
	(0.472)	(0.439)	(0.701)	(0.812)	(0.207)	(0.186)
IQA x CCR	-0.057	-0.035	-0.024	0.018	0.020	0.026
	(0.167)	(0.179)	(0.371)	(0.324)	(0.281)	(0.166)
CCR	-0.052	-0.028	-0.018	**-0.091*** **	-0.040	-0.044
	(0.064)	(0.455)	(0.353)	(0.000)	(0.301)	(0.107)
Year Slope (pre CCR)		0.005			-0.000	
		(0.739)			(0.996)	
Year Slope (CCR)		0.014			-0.023	
		(0.422)			(0.237)	
Cohort Controls	X	X	X	X	X	X
Intercept	**-0.018***	-0.005	0.006	**0.038****	0.023	0.014
	(0.034)	(0.856)	(0.714)	(0.005)	(0.316)	(0.470)
N	218	218	218	194	194	194
AIC	-469.5	-192.6	-176.3	-589.7	-243.6	-194.3
BIC	-459.3	-168.9	-159.4	-576.6	-210.9	-168.2
IQA Composite in CCR Years (IQA coefficient + CCR interaction Term)	-0.032	-0.021	-0.017	0.022	**0.036***	**0.043****
	(0.165)	(0.268)	(0.377)	(0.275)	(0.015)	(0.004)

APPENDIX G

Chetty, Friedman, & Rockoff Methodology

Chetty, Friedman, and Rockoff (2014) apply a value-added methodology which they describe as differentiated from a number of other models which "typically assume that each teacher's quality is fixed over time and thus place equal weight on test scores in all classes taught by the teacher when forecasting teacher quality" despite the fact that "test scores from more recent classes are better predictors of current teacher quality, indicating that teacher quality fluctuates over time" (p.2). However, the approach of this analysis, which estimates teacher value-added utilizing separate regressions for each district and year, does not rely on an assumption that teacher quality or effectiveness is fixed over time.

At the same time, the approach employed by Chetty and colleagues also contains features which limit year-to-year fluctuations in teacher value-added estimates. When estimating teacher value-added for teacher *j* in year *t*, their approach utilizes student tests scores from students assigned to teacher *i* in all years except year *t* (i.e., years t-2, t-1, t+1, t+2), in a leave-year-out/jackknife approach also utilized in Jacob, Lefgren, and Sims (2010). This approach is justified by these authors in that it excludes teacher-by-year (η_{jt}) factors which effect classroom performance: Jacobs and colleagues provide examples such as "a test administered on a hot day or unusually good match quality between the teacher and students" as factors which might fit into this category (p. 928). These issues are likely more of a concern in analyses where elementary-level data is used, where only one class is nested within teacher year such that any idiosyncratic shock for class *k* in year *t* like those given as examples by Jacob and colleagues (say η_{jkt}) cannot be easily separated from a year-specific component of teacher contributions to teacher learning (say τ_{jt}). In circumstances where teachers teach more than one section in a year, as in a middle-school context, these are easier to separate, with some expectation that on average, the class-level idiosyncratic shocks will tend to cancel each other out within teacher-year. In addition, this class-by-year error component can be expected to be much larger in models without class-level controls for demographics and average prior achievement (as in Jacob et al,

2010), compared to models where these controls are included, as in Chetty and colleagues (2014) or the MET Project methodology emulated here.

In addition, this jackknife approach to estimating teacher value-added was also necessitated by the research question and estimation methods of these two analyses. In Jacob and colleagues' analysis, teacher value-added in year *t* for teacher *j* (θ_{jt}) was estimated without data from student scores in year *t* because these estimates were then used in a model in which the score for student *i* assigned to teacher *j* in year *t* was regressed on the student's score in year *t-1* along with the teacher value-added estimate for that year (θ_{tj}), in order to estimate persistence of teacher value-add. Similarly, Chetty and colleagues use their value added estimates for teacher *j* in year *t* – estimated using data for all of teacher *j*'s students in all years not *t* – as a regressor in an equation predicting the student achievement for teacher *j*'s students in year *t* in order to measure the amount of forecast bias present in these estimates. In short, the estimation of a teacher effect *j* in year *t* which drew from data for teacher *j*'s student in any year except *t* was necessitated by the research question in each of these analyses.

The research question of the current analysis suggests that value-added approaches which smooth or attenuate year-to-year fluctuations of teacher value-added estimates my not be the most appropriate choice for this research question. Because this research investigates the influence of a change in content standards and assessments which occurs in the second half of this longitudinal data, we might expect to see a discontinuity or step-wise function which might be diminished using measures with smooth estimates over time

Perhaps the largest practical disadvantage in applying this approach to these data is that it significantly reduces the sample size. Both the Jacobs and colleagues analysis and Chetty and colleagues analysis confirms that this estimation approach cannot estimate value-added for teachers who only appear during one year in the data set. For the data from this analysis, that would mean reduction of sample size of 34% (see Table F.1).

Table G1. Percentage of teacher-year cases with only one year of data, by district.

	Dist B (Years 1-7)		Dist D (Years 1-7)		Both Districts (Y1-7)	
Teacher-year cases for teachers with only 1 year of data	68	30.4%	79	37.6%	147	33.9%
Teacher-year cases for teachers with more than 1 year of data	156	69.6%	131	62.4%	287	66.1%
	224	**100.0%**	**210**	**100.0%**	**434**	**100%**

Table G2 Percentage of observations per teacher, by district

	District B		District D	
Observations	Count	Percent	Count	Percent
1	68	57.1%	79	63.2%
2	25	21.0%	28	22.4%
3	9	7.6%	7	5.6%
4	8	6.7%	6	4.8%
5	7	5.9%	2	1.6%
6	2	1.7%	1	0.8%
7	0	0.0%	2	1.6%
	119	**100.0%**	**125**	**100.0%**

APPENDIX H

Sensitivity Analyses: Varying the Cutoff for Class-size Exclusion

All things being equal, the precision of the teacher-value estimate for a given teacher is a function of the number of students associated with that teacher: the greater the number of students which can be associated with a given teacher, the less uncertain the estimates of the teacher's average contributions to student learning from one testing cycle to another. Quantitative researchers have dealt with these limitation in a number of ways. Some have employed shrinkage weights to teachers' value-added estimates which take into account both the numbers of students assigned to a teacher and the variation in growth scores between those students: for teachers with smaller classes or who have larger variability in measured learning gains, their estimates are less precise and will be weighted to fall closer to the mean (Herrmann, Walsh, & Eisenberg, 2016). This strategy allows for value-added to be estimated for all teachers, while also serving to produce more conservative estimates for school accountability systems where estimated value-added may be linked with consequences: teachers with fewer students and very high or very low value-added estimates will see their value-added estimates move closer to that of the average teacher after shrinkage weights have been applied.

Another strategy to deal with the lack of precision associated with smaller numbers of students assigned to some teachers is to apply a cutoff. A number of analyses which estimate teacher value-added employ a cutoff for the minimum number of students that can be associated with a teacher in order for the teacher to be included in the analytical sample. A brief investigation of some of the literature reveals a couple of analyses which use 10 students per teachers as a minimum for inclusion in the analytical sample (Kane & Staiger, 2008; Sass, Semykina, & Harris, 2014), with one using a cutoff of 15 students per teacher (Isenberg & Hock, 2010).

For the present analysis, a larger cutoff of a minimum of 30 students per teacher-year was used. This excluded less than 8 percent of the total teacher-year observations which would be included in a bare-minimum threshold of 5 students per teacher-year (See Table G1.). This compares to a excluding less than 1 percent at the 10 students per teacher-year threshold, or less than 3 percent of teachers at the 15 student per teacher threshold.

The choice of a minimum student-per-teacher-year threshold does in some cases influence both the value-added to IQA regression coefficient as well as the precision of that estimates (i.e., the standard errors), though the degree of influence varies by district, test period, and estimation method (see Figure G3).

In District B, changes in the minimum student-per-teacher-year threshold do not seem to influence the point estimate, with the exception of the point estimates from the teacher fixed-effects estimates, which generally seem to decrease – becoming more negative – as the threshold increases. In District D, the point estimates for the IQA regression coefficients increase as the threshold is increased from 15 to 20 student; this is more marked in the IQA coefficients for the CCR period.

In general, increasing the threshold number does not seem to change substantially the precision of the estimate, with the exception of the coefficient for IQA in District D in the CCR years where, as the threshold is increased from 15 to 20 students, the 95 percent confidence intervals on the IQA coefficient in CCR years decrease by 30 percent in the teacher fixed-effects model and decrease by 20 percent in the POLS models.

Figure H1a-b. Histogram distribution of number of student contributing to the value-added estimate in each teacher-year observation

Figure H1a. District B.

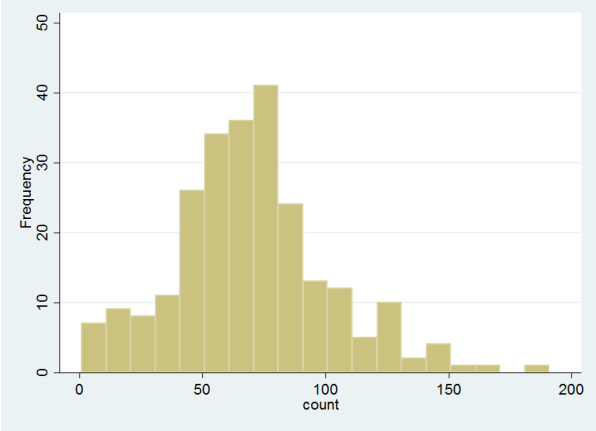


Figure H1b. District D.

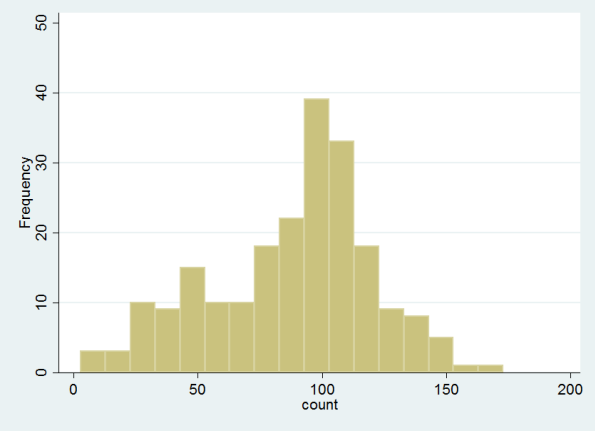


Figure H2. Decline in teacher-year sample size as cutoff for inclusion (by number of students contributing to value-added estimate) changes.

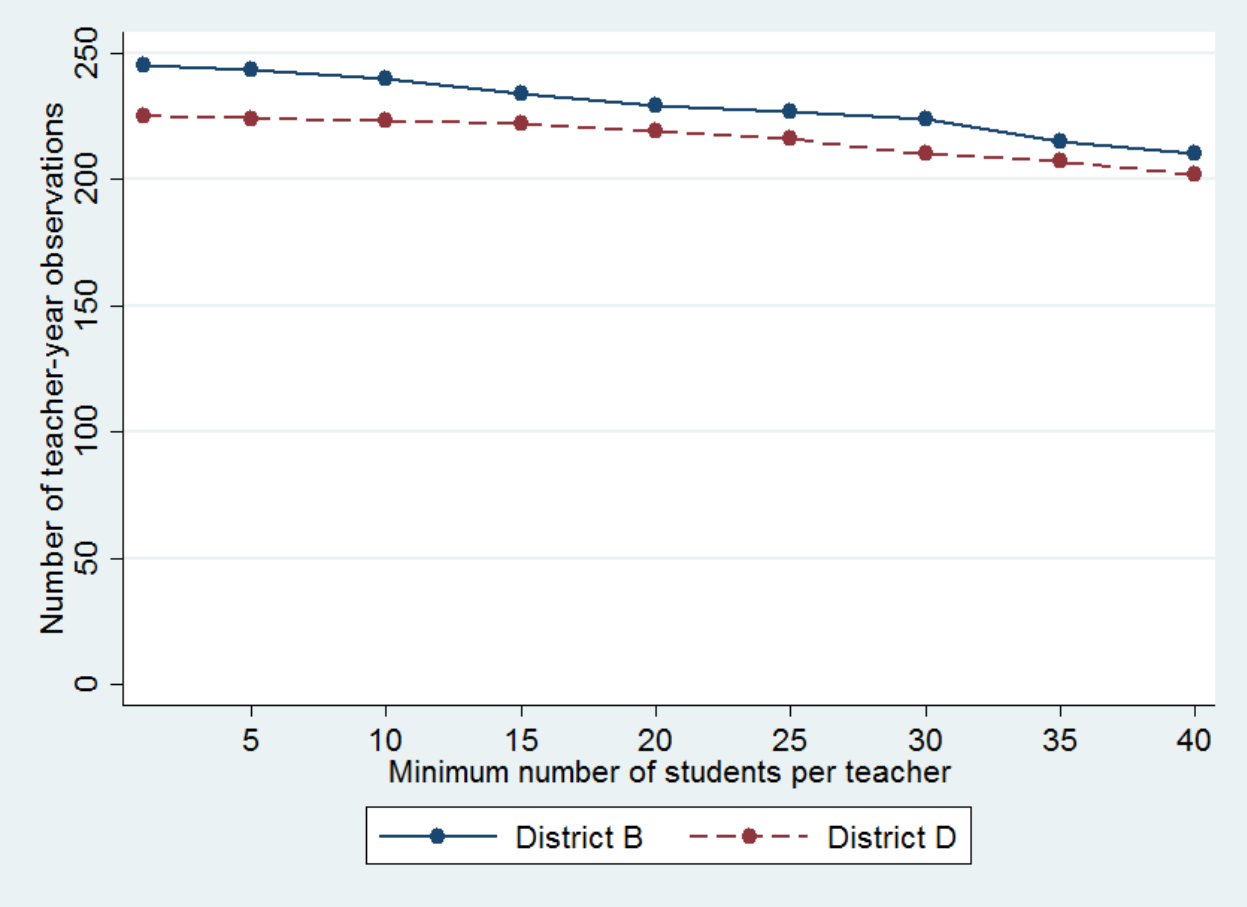
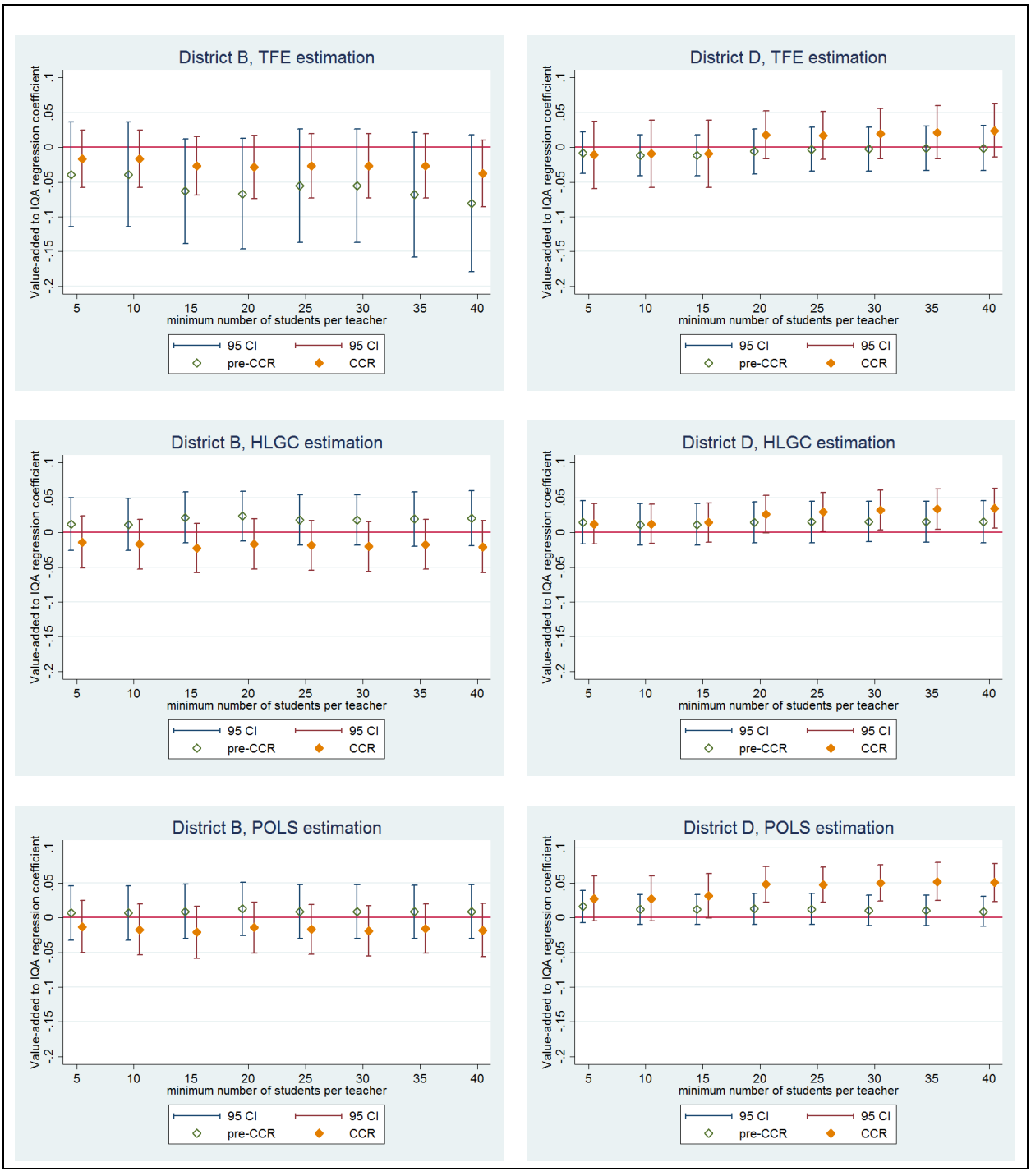


Table H1. Decline in teacher-year sample size as cutoff for inclusion (by number of students contributing to value-added estimate) changes.

	Dist B		Dist D	
Minimum students per teacher-year	Teacher-year obs	% excluded from 5 minimum obs	Teacher-year obs	% excluded from 5 minimum obs
No min.	245	0.0%	225	0.0%
5	243	0.8%	224	0.4%
10	240	2.0%	223	0.9%
15	234	4.5%	222	1.4%
20	229	6.5%	219	2.7%
25	227	7.3%	216	4.2%
30	224	8.6%	210	7.1%
35	215	12.2%	207	8.7%
40	210	14.3%	202	11.4%

Figure H3a-f. Change in point estimates, precision of point estimates as cutoff (minimum number of students contributing to teacher-year value-added estimate) changes, by district, test regime, and estimation method.



APPENDIX I

Pre-CCR & CCR Grades 6-8 Math Standards in Districts B & D:Details

Table I1. Pre-CCR & CCR Grade 6 Math Standards, Districts B

Standards Set (# of standards)	**Standards**
Pre-CCR, District B, Grade 6 (38 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 1, Standard D; Domain 1, Standard E; Domain 1, Standard F; Domain 2, Standard A; Domain 2, Standard B; Domain 2, Standard C; Domain 2, Standard D; Domain 2, Standard E; Domain 3, Standard A; Domain 3, Standard B; Domain 3, Standard C; Domain 4, Standard A; Domain 4, Standard B; Domain 5, Standard (NULL); Domain 6, Standard A; Domain 6, Standard B; Domain 7, Standard (NULL); Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 8, Standard D; Domain 9, Standard A; Domain 9, Standard B; Domain 10, Standard A; Domain 10, Standard B; Domain 10, Standard C; Domain 10, Standard D; Domain 11, Standard A; Domain 11, Standard B; Domain 11, Standard C; Domain 11, Standard D; Domain 12, Standard A; Domain 12, Standard B; Domain 13, Standard A; Domain 13, Standard B;
CCR, District B, Grade 6 (59 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 1, Standard D; Domain 1, Standard E; Domain 1, Standard F; Domain 1, Standard G; Domain 2, Standard A; Domain 2, Standard B; Domain 2, Standard C; Domain 2, Standard D; Domain 2, Standard E; Domain 3, Standard A; Domain 3, Standard B; Domain 3, Standard C; Domain 3, Standard D; Domain 3, Standard E; Domain 4, Standard A; Domain 4, Standard B; Domain 4, Standard C; Domain 4, Standard D; Domain 4, Standard E; Domain 4, Standard F; Domain 4, Standard G; Domain 4, Standard H; Domain 5, Standard A; Domain 5, Standard B; Domain 5, Standard C; Domain 6, Standard A; Domain 6, Standard B; Domain 6, Standard C; Domain 7, Standard A; Domain 7, Standard B; Domain 7, Standard C; Domain 7, Standard D; Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 8, Standard D; Domain 9, Standard A; Domain 9, Standard B; Domain 9, Standard C; Domain 10, Standard A; Domain 10, Standard B; Domain 11, Standard (NULL); Domain 12, Standard A; Domain 12, Standard B; Domain 12, Standard C; Domain 12, Standard D; Domain 13, Standard A; Domain 13, Standard B; Domain 14, Standard A; Domain 14, Standard B; Domain 14, Standard C; Domain 14, Standard D; Domain 14, Standard E; Domain 14, Standard F; Domain 14, Standard G; Domain 14, Standard H;

Table I2. Pre-CCR & CCR Grade 7 Math Standards, Districts B

Standards Set (# of standards)	Standards
Pre-CCR, District B, Grade 7 (43 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 2, Standard A; Domain 2, Standard B; Domain 2, Standard C; Domain 2, Standard D; Domain 2, Standard E; Domain 2, Standard F; Domain 2, Standard G; Domain 3, Standard A; Domain 3, Standard B; Domain 4, Standard A; Domain 4, Standard B; Domain 4, Standard C; Domain 5, Standard A; Domain 5, Standard B; Domain 6, Standard A; Domain 6, Standard B; Domain 6, Standard C; Domain 6, Standard D; Domain 7, Standard A; Domain 7, Standard B; Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 9, Standard A; Domain 9, Standard B; Domain 9, Standard C; Domain 10, Standard A; Domain 10, Standard B; Domain 11, Standard A; Domain 11, Standard B; Domain 12, Standard A; Domain 12, Standard B; Domain 13, Standard A; Domain 13, Standard B; Domain 13, Standard C; Domain 13, Standard D; Domain 14, Standard A; Domain 14, Standard B; Domain 15, Standard A; Domain 15, Standard B;
CCR, District B, Grade 7 (50 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 1, Standard D; Domain 1, Standard E; Domain 1, Standard F; Domain 1, Standard G; Domain 2, Standard (NULL); Domain 3, Standard A; Domain 3, Standard B; Domain 4, Standard A; Domain 4, Standard B; Domain 4, Standard C; Domain 4, Standard D; Domain 4, Standard E; Domain 5, Standard A; Domain 5, Standard B; Domain 5, Standard C; Domain 6, Standard A; Domain 6, Standard B; Domain 6, Standard C; Domain 6, Standard D; Domain 6, Standard E; Domain 6, Standard F; Domain 6, Standard G; Domain 6, Standard H; Domain 6, Standard I; Domain 7, Standard (NULL); Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 9, Standard A; Domain 9, Standard B; Domain 9, Standard C; Domain 9, Standard D; Domain 10, Standard A; Domain 10, Standard B; Domain 10, Standard C; Domain 11, Standard A; Domain 11, Standard B; Domain 11, Standard C; Domain 12, Standard A; Domain 12, Standard B; Domain 12, Standard C; Domain 13, Standard A; Domain 13, Standard B; Domain 13, Standard C; Domain 13, Standard D; Domain 13, Standard E; Domain 13, Standard F;

Table I3. Pre-CCR & CCR Grade 8 Math Standards, Districts B

Standards Set (# of standards)	Standards
Pre-CCR, District B, Grade 8 (43 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 1, Standard D; Domain 1, Standard E; Domain 2, Standard A; Domain 2, Standard B; Domain 2, Standard C; Domain 2, Standard D; Domain 3, Standard A; Domain 3, Standard B; Domain 4, Standard (NULL); Domain 5, Standard A; Domain 5, Standard B; Domain 6, Standard A; Domain 6, Standard B; Domain 7, Standard A; Domain 7, Standard B; Domain 7, Standard C; Domain 7, Standard D; Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 9, Standard A; Domain 9, Standard B; Domain 10, Standard A; Domain 10, Standard B; Domain 11, Standard A; Domain 11, Standard B; Domain 11, Standard C; Domain 12, Standard A; Domain 12, Standard B; Domain 12, Standard C; Domain 13, Standard A; Domain 13, Standard B; Domain 14, Standard A; Domain 14, Standard B; Domain 14, Standard C; Domain 14, Standard D; Domain 15, Standard A; Domain 15, Standard B; Domain 16, Standard A; Domain 16, Standard B
CCR, District B, Grade 8 (52 Standards)	Domain 1, Standard A; Domain 1, Standard B; Domain 1, Standard C; Domain 1, Standard D; Domain 1, Standard E; Domain 1, Standard F; Domain 1, Standard G; Domain 2, Standard A; Domain 2, Standard B; Domain 2, Standard C; Domain 2, Standard D; Domain 3, Standard A; Domain 3, Standard B; Domain 3, Standard C; Domain 4, Standard A; Domain 4, Standard B; Domain 4, Standard C; Domain 5, Standard A; Domain 5, Standard B; Domain 5, Standard C; Domain 5, Standard D; Domain 5, Standard E; Domain 5, Standard F; Domain 5, Standard G; Domain 5, Standard H; Domain 5, Standard I; Domain 6, Standard A; Domain 6, Standard B; Domain 6, Standard C; Domain 7, Standard A; Domain 7, Standard B; Domain 7, Standard C; Domain 7, Standard D; Domain 8, Standard A; Domain 8, Standard B; Domain 8, Standard C; Domain 8, Standard D; Domain 9, Standard (NULL); Domain 10, Standard A; Domain 10, Standard B; Domain 10, Standard C; Domain 10, Standard D; Domain 11, Standard A; Domain 11, Standard B; Domain 11, Standard C; Domain 12, Standard A; Domain 12, Standard B; Domain 12, Standard C; Domain 12, Standard D; Domain 12, Standard E; Domain 12, Standard F; Domain 12, Standard G

Table I4. Pre-CCR & CCR Grades 6-7 Math Standards, Districts D

Standards Set (# of standards)	Standards
Pre-CCR, District D, Grade 6 (37 Standards)	Algebra:1.1; Algebra:1.2; Algebra:1.3; Algebra:1.4; Algebra:1.5; Algebra:2.1; Algebra:2.2; Algebra:3.1; Data Analysis & Prob:1.1; Data Analysis & Prob:1.2; Data Analysis & Prob:1.4; Data Analysis & Prob:2.1; Data Analysis & Prob:4.1; Data Analysis & Prob:4.2; Data Analysis & Prob:4.3; Geometry:1.1; Geometry:1.2; Geometry:1.3; Geometry:1.4; Geometry:1.5; Geometry:2.1; Geometry:2.2; Geometry:2.3; Geometry:3.1; Measurement:1.1; Measurement:1.2; Measurement:1.3; Measurement:2.1; Num Prop & Op:1.1; Num Prop & Op:1.2; Num Prop & Op:1.3; Num Prop & Op:2.1; Num Prop & Op:3.1; Num Prop & Op:3.2; Num Prop & Op:4.1; Num Prop & Op:5.1; Num Prop & Op:5.2
CCR, District D, Grade 6 (29 Standards)	Exp & Equat:A.1; Exp & Equat:A.2; Exp & Equat:A.3; Exp & Equat:A.4; Exp & Equat:B.5; Exp & Equat:B.6; Exp & Equat:B.7; Exp & Equat:B.8; Exp & Equat:C.9; Geo:A.1; Geo:A.2; Geo:A.3; Geo:A.4; Number Sys:A.1; Number Sys:B.2; Number Sys:B.3; Number Sys:B.4; Number Sys:C.5; Number Sys:C.6; Number Sys:C.7; Number Sys:C.8; Ratios & Prop:A.1; Ratios & Prop:A.2; Ratios & Prop:A.3; Stat & Prob:A.1; Stat & Prob:A.2; Stat & Prob:A.3; Stat & Prob:B.4; Stat & Prob:B.5;
Pre-CCR, District D, Grade 7 (38 Standards)	Algebra:1.1; Algebra:1.2; Algebra:1.3; Algebra:1.5; Algebra:2.1; Algebra:2.2; Algebra:3.1; Data Analysis & Prob:1.1; Data Analysis & Prob:1.2; Data Analysis & Prob:1.3; Data Analysis & Prob:1.4; Data Analysis & Prob:1.5; Data Analysis & Prob:2.1; Data Analysis & Prob:4.1; Data Analysis & Prob:4.2; Data Analysis & Prob:4.3; Geometry:1.1; Geometry:1.2; Geometry:1.3; Geometry:1.4; Geometry:2.2; Geometry:2.3; Geometry:3.1; Measurement:1.1; Measurement:1.2; Measurement:1.3; Measurement:1.4; Measurement:2.1; Num Prop & Op:1.1; Num Prop & Op:1.2; Num Prop & Op:1.3; Num Prop & Op:2.1; Num Prop & Op:3.1; Num Prop & Op:3.2; Num Prop & Op:3.3; Num Prop & Op:4.1; Num Prop & Op:5.1; Num Prop & Op:5.2
CCR, District D, Grade 7 (24 Standards)	Exp & Equat:A.1; Exp & Equat:A.2; Exp & Equat:B.3; Exp & Equat:B.4; Geometry:A.1; Geometry:A.2; Geometry:A.3; Geometry:B.4; Geometry:B.5; Geometry:B.6; Number Sys:A.1; Number Sys:A.2; Number Sys:A.3; Ratios & Prop:A.1; Ratios & Prop:A.2; Ratios & Prop:A.3; Stats & Prob:A.1; Stats & Prob:A.2; Stats & Prob:B.3; Stats & Prob:B.4; Stats & Prob:C.5; Stats & Prob:C.6; Stats & Prob:C.7; Stats & Prob:C.8

Table I5. Pre-CCR & CCR Grade 8 Math Standards, Districts D

Standards Set (# of standards)	**Standards**
Pre-CCR, District D, Grade 8 (39 Standards)	Algebra:1.1; Algebra:1.2; Algebra:1.5; Algebra:2.1; Algebra:2.2; Algebra:3.1; Data Analysis & Prob:1.1; Data Analysis & Prob:1.2; Data Analysis & Prob:1.4; Data Analysis & Prob:1.5; Data Analysis & Prob:2.1; Data Analysis & Prob:3.1; Data Analysis & Prob:4.1; Data Analysis & Prob:4.2; Data Analysis & Prob:4.3; Data Analysis & Prob:4.4; Geometry:1.1; Geometry:1.2; Geometry:1.3; Geometry:1.4; Geometry:2.1; Geometry:2.2; Geometry:2.3; Geometry:3.1; Measurement:1.1; Measurement:1.2; Measurement:1.3; Measurement:1.4; Measurement:1.5; Measurement:1.6; Measurement:2.1; Number Sense:1.1; Number Sense:1.2; Number Sense:1.3; Number Sense:2.1; Number Sense:3.1; Number Sense:3.2; Number Sense:4.1; Number Sense:5.2
CCR, District D, Grade 8 (28 Standards)	Exp & Equat:A.1; Exp & Equat:A.2; Exp & Equat:A.3; Exp & Equat:A.4; Exp & Equat:B.5; Exp & Equat:B.6; Exp & Equat:C.7; Exp & Equat:C.8; Functions:A.1; Functions:A.2; Functions:A.3; Functions:B.4; Functions:B.5; Geometry:A.1; Geometry:A.2; Geometry:A.3; Geometry:A.4; Geometry:A.5; Geometry:B.6; Geometry:B.7; Geometry:B.8; Geometry:C.9; Number Sys:A.1; Number Sys:A.2; Stat & Prob:A.1; Stat & Prob:A.2; Stat & Prob:A.3; Stat & Prob:A.4

APPENDIX J

Regression Approaches Addressing Measurement Error in Prior Year Test Scores

This appendix describes regression analysis of these teacher-linked student test score data following a method used to address measurement error in students' prior year test scores in relatively recent literature (e.g. Herrmann, Walsh, & Isenberg, 2016). In this approach, the first-stage regression is modeled as follows:

$$Y_{tiy} = \lambda_y Y_{i(y-1)} + \alpha \boldsymbol{X}_i + \varepsilon_{tiy} \quad [J.1]$$

where the observed achievement score *Y* for student *i* associated with teacher *t* in year *y* is modeled as a function of the student's test score in year *y-1*. $\boldsymbol{X}$ denotes a vector of student demographic variables functioning as control variables – grade, gender, race/ethnicity, special education status, and English learner status. In this first step, I estimate Equation J.1 adjusting for measurement error in the pre-test using errors-in-variables correction (eivreg in Stata). These first-stage regressions are estimated separately by district for each year of data. I then recover the measurement-error corrected estimates for the pretest coefficient ($\hat{\lambda}_y$) to calculate a residual equal to the observed test score minus the effect or influence of the pre-test score, as in Equation (I.2):

$$\hat{G}_{tiy} = Y_{tiy} - \hat{\lambda}_y Y_{i(y-1)} \quad [J.2]$$

This calculated term $\hat{G}_{tiy}$ is then used as the dependent variable in a second stage regression:

$$\hat{G}_{tiy} = \alpha' \boldsymbol{X}_i + \eta \boldsymbol{T}_{ty} + \varepsilon_{tiy} \quad [J.3]$$

where the term $\hat{G}_{tiy}$ is modeled as a function of student demographics (vector $\boldsymbol{X}$), an average teacher-level fixed effect for teacher *t* in year *y* (denoted by a vector of teacher-by-year specific indicator variables $\boldsymbol{T}$), and an error term (ε_{tiy}), as in Equation J.3.

I located state documents providing alpha reliabilities for a sub-sample of the District B and D assessments in grades 6 through 8 across the seven years encompassing this study (Table J.1). These data suggest that reliabilities – at least by this measure – are centered around 0.90, with the minimum and maximum alpha reliabilities for this set of tests at 0.882 and 0.932, respectively (Table J.1). Because information for each district, grade, and year in question was not available – and because these regressions results serve as a robustness check, I performed a series of robustness check regressions corresponding to scenarios where the reliability was constant across time and districts, while also exploring the difference in estimates and results as the reliability was allowed to vary between 0.75 and 1.00 across different simulations. This range of reliabilities is comparable to the range of reliabilities explored by Lockwood, McCaffrey, and Savage (2017) in their evaluation of the implementation of eivreg for estimating teacher value-added in the presence of measurement error in the pre-test.

The relationship between these value-added estimates and the IQA measure of classroom instructions were estimated using a fully-interacted pooled ordinary least squares (POLS) regression and compared with results from the dissertation's main model (i.e. the MET methodology value-added estimates). The POLS model was chosen for the purposes of this comparison because this approach produced the highest number of statistically significant differences between IQA and teacher value-added, both within and between districts (see Appendix D) . Specifically, the fully-interacted POLS analysis of the MET methodology value-added estimates show four different linear combinations of coefficients which are significant at conventional levels ($p<0.05$).

Results from the regressions which make adjustments for measurement error show that two of these four linear combinations of variables remain consistently significant at conventional levels across most of the range of reliability values explored, including the 0.90 level of reliability which most closely approximates the reported alpha reliabilities of these assessments (Table I4). These two coefficients represent (1) IQA composite predicting teacher value-added in District D in the CCR years, along with (2) the difference between the IQA composite coefficients in District B and District D in the CCR years). This robustness analysis contributes some evidence that these are some of the more robust significant relationships from these data, given that these two linear combinations of coefficients were also significant at conventional

levels in the results from the hierarchical linear growth curve (HLGC) regressions (see Appendix E, Table E2).

Table J1. Pairwise correlation coefficients from value-added estimates which address measurement error in the pre-test, across a range of potential test reliabilities,

	VAM, α=0.75	VAM, α=0.80	VAM, α=0.85	VAM, α=0.90	VAM, α=0.95	VAM, α=1.00	MET Methodology
VAM, α=0.75	1.00						
VAM, α=0.80	0.97	1.00					
VAM, α=0.85	0.90	0.98	1.00				
VAM, α=0.90	0.79	0.91	0.98	1.00			
VAM, α=0.95	0.68	0.83	0.93	0.99	1.00		
VAM, α=1.00	0.58	0.75	0.88	0.95	0.99	1.00	
MET Methodology	0.67	0.77	0.83	0.85	0.84	0.81	1.00

Table J2 Alpha reliabilities of state mathematics exams in a sample of exams across districts, middle-grades, and study years.

Dist	Study Year	Calendar Year	Standards	Grade	alpha (avg. where multiple forms)
B	3	2010	pre-CCR	6	0.909
B	3	2010	pre-CCR	7	0.904
B	3	2010	pre-CCR	8	0.907
B	4	2011	pre-CCR	6	0.908
B	4	2011	pre-CCR	7	0.904
B	4	2011	pre-CCR	8	0.906
B	5	2012	CCR	6	0.932
B	5	2012	CCR	7	0.915
B	5	2012	CCR	8	0.884
B	6	2013	CCR	6	0.932
B	6	2013	CCR	7	0.908
B	6	2013	CCR	8	0.895
B	7	2014	CCR	6	0.926
B	7	2014	CCR	7	0.916
B	7	2014	CCR	8	0.911
D	2	2009	pre-CCR	6	0.890
D	2	2009	pre-CCR	7	0.890
D	2	2009	pre-CCR	8	0.890
D	4	2011	pre-CCR	6	0.882
D	4	2011	pre-CCR	7	0.892
D	4	2011	pre-CCR	8	0.886
D	5	2012	CCR	6	0.900
D	5	2012	CCR	7	0.910
D	5	2012	CCR	8	0.900
D	6	2013	CCR	6	0.890
D	6	2013	CCR	7	0.890
D	6	2013	CCR	8	0.890
D	7	2014	CCR	6	0.910
D	7	2014	CCR	7	0.910
D	7	2014	CCR	8	0.900
			Min	Max	Average
B pre-CCR			0.904	0.909	0.906
B CCR			0.884	0.932	0.913
D pre-CCR			0.882	0.892	0.888
D CCR			0.890	0.910	0.900

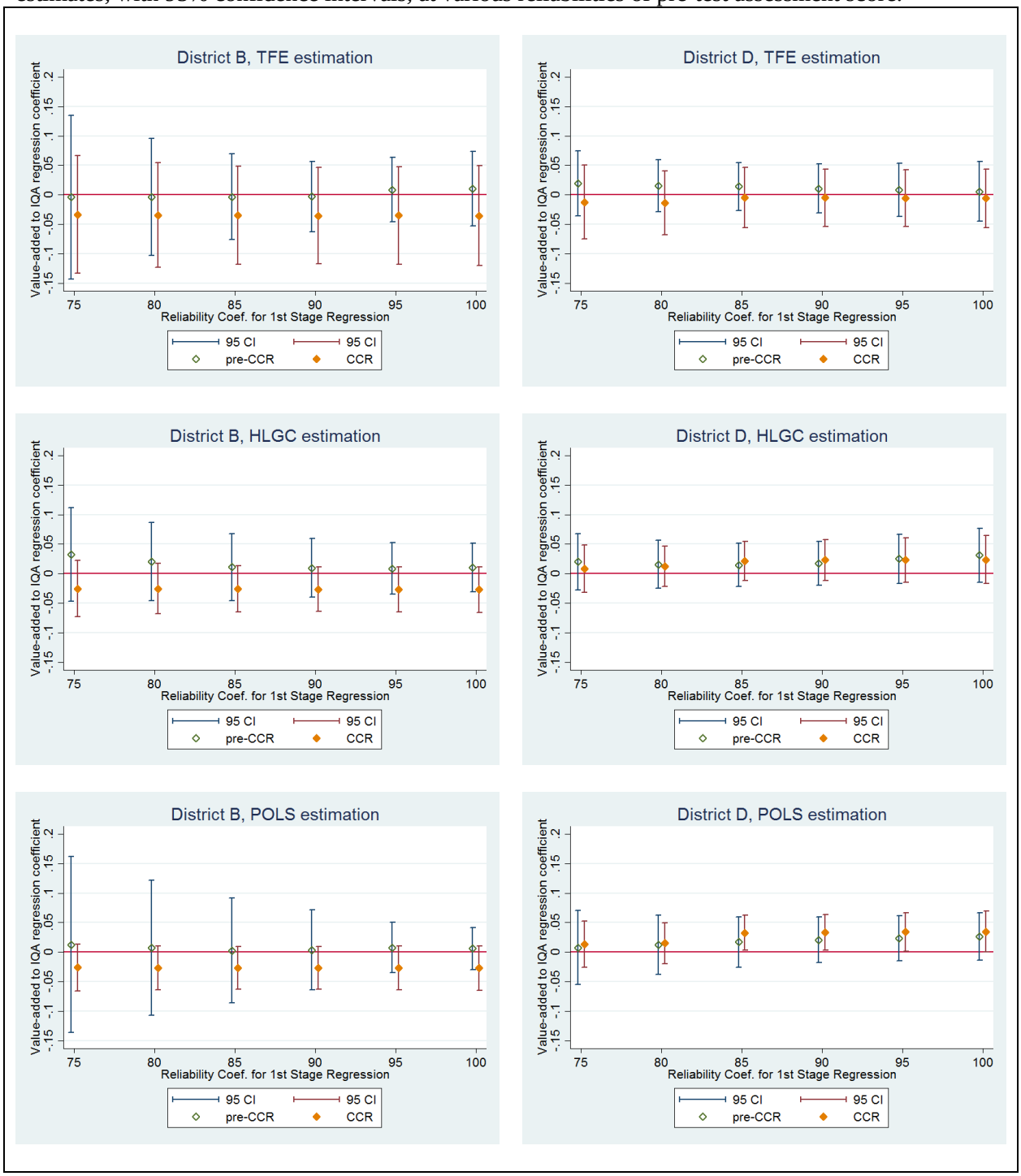
Table J3 Fully-interacted pooled ordinary least squares (POLS) regression table of teacher value-added estimates using two-step eivreg procedure. With reliabilities ranging from 0.75 to 1.00, and including results from main model (MET methodology). Standard errors are cluster-adjusted at the teacher level. (P-values in parentheses)

	Results from 2-step errors-in-variables regressions						
	ρ=0.75	ρ=0.80	ρ=0.85	ρ=0.90	ρ=0.95	ρ=1.00	MET
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IQA Composite	0.013	0.007	0.003	0.004	0.008	0.006	0.007
	(0.867)	(0.902)	(0.953)	(0.917)	(0.723)	(0.741)	(0.705)
IQAxCCR	-0.039	-0.034	-0.029	-0.031	-0.035	-0.033	-0.025
	(0.629)	(0.593)	(0.557)	(0.449)	(0.224)	(0.208)	(0.343)
CCR	0.002	-0.013	-0.024	-0.036	-0.040	**-0.046+**	-0.019
	(0.979)	(0.791)	(0.521)	(0.238)	(0.115)	(0.052)	(0.338)
Dist. D	0.099	0.086	0.071	0.061	0.059	0.054	0.029
	(0.321)	(0.348)	(0.432)	(0.515)	(0.546)	(0.598)	(0.594)
Dist. DxIQA	-0.005	0.005	0.014	0.017	0.016	0.020	0.004
	(0.948)	(0.935)	(0.776)	(0.672)	(0.583)	(0.454)	(0.874)
Dist. DxCCR	-0.035	-0.042	-0.040	-0.042	-0.050	-0.054	-0.013
	(0.628)	(0.475)	(0.428)	(0.358)	(0.247)	(0.218)	(0.704)
Dist. DxCCRxIQA	0.045	0.037	0.045	0.043	0.045	0.042	**0.062***
	(0.610)	(0.602)	(0.424)	(0.361)	(0.240)	(0.274)	(0.048)
Cohort Fixed Effects	X	X	X	X	X	X	X
Intercept	-0.006	-0.004	-0.003	0.001	-0.003	-0.003	0.006
	(0.920)	(0.930)	(0.944)	(0.984)	(0.903)	(0.876)	(0.734)
434	434	434	435	435	435	434	434
156.020	7.182	-98.173	-152.797	-172.162	-154.968	156.020	-379.6
208.969	60.132	-45.223	-99.818	-119.183	-101.988	208.969	-326.6

Table J4 Linear combinations of coefficients from fully-interacted pooled ordinary least squares (POLS) regression table of teacher value-added estimates using two-step eivreg procedure. With reliabilities ranging from 0.75 to 1.00, and including results from main model (MET methodology). Standard errors are cluster-adjusted at the teacher level. (P-values in parentheses)

	Results from 2-step errors-in-variables regressions						
	ρ=0.75	ρ=0.80	ρ=0.85	ρ=0.90	ρ=0.95	ρ=1.00	MET
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
IQA Composite in CCR Years	-0.026	-0.027	-0.027	-0.027	-0.027	-0.027	-0.017
(Dist B)	(0.199)	(0.163)	(0.149)	(0.147)	(0.153)	(0.162)	(0.330)
IQA Composite in pre-CCR	0.007	0.012	0.017	0.021	0.024	0.026	0.011
Years (Dist D)	(0.816)	(0.627)	(0.435)	(0.297)	(0.223)	(0.194)	(0.330)
IQA Composite in CCR Years	0.013	0.015	**0.032***	**0.033***	**0.034***	**0.035+**	**0.048*****
(Dist D)	(0.498)	(0.388)	(0.032)	(0.032)	(0.040)	(0.051)	(0.000)
Difference :IQA Coef. in Dist.	-0.005	0.005	0.014	0.017	0.016	0.020	0.004
D and B (Pre –CCR Years)	(0.948)	(0.935)	(0.776)	(0.6782)	(0.583)	(0.454)	(0.874)
Difference :IQA Coef. in Dist.	0.040	0.042	**0.059***	**0.060***	**0.061***	**0.062***	**0.066****
D and B (CCR Years)	(0.164)	(0.108)	(0.014)	(0.013)	(0.015)	(0.019)	(0.003)
Difference: Pre-CCR to CCR	-0.039	-0.034	-0.029	-0.031	-0.035	-0.033	-0.025
Dist B	(0.629)	(0.593)	(0.557)	(0.449)	(0.224)	(0.208)	(0.343)
Difference: Pre-CCR to CCR	0.006	0.003	0.016	0.013	0.010	0.008	**0.037***
Dist D	(0.866)	(0.927)	(0.545)	(0.607)	(0.684)	(0.759)	(0.032)
Difference :IQA "slopes". in	0.045	0.037	0.045	0.043	0.045	0.042	**0.062***
Dist D and B (ΔD-ΔB)	(0.610)	(0.602)	(0.424)	(0.361)	(0.240)	(0.274)	(0.048)

Figure J1a-f. Point estimates of relationship between IQA composite and teacher value-added estimates, with 95% confidence intervals, at various reliabilities of pre-test assessment score.



APPENDIX K

A Description of the Influence of Measurement Error on Estimates Using Simulated Data

In order to investigate the consequences of measurement error in both (a) the prior achievement scores at the student level and (b) the IQA composite scores derived during observation of teachers' classrooms, simulated data was created to emulate some of the characteristics of the data used in this analysis. The simulated data set was generated using the following relationships gleaned from multilevel regression of data from District D in Year 7:

$$Achievement_{it} = (-0.09)FRL_{it} + (0.0)\overline{FRL}_{jkt} + (0.62)Achievement_{true_{it-1}} + (0.31)\overline{Achievement_{true}}_{jkt-1} + \tau_{jt} + u_{it} \quad [K1]$$

$$\tau_{jt} = 0.03IQA_{true_{jt}} + \sqrt{(0.16^2 - 0.03^2)}\, u'_{jt} \quad [K2]$$

$$\tau_{jt} \sim N(0.0, 0.16) \quad [K3]$$

$$\varepsilon_{it} \sim N(0.0, 0.54) \quad [K4]$$

$$\rho\left(FRL, Achievement_{true_{t-1}}\right) = -0.44 \quad [K5]$$

Unless noted above, predictor variables are uncorrelated and have a mean of zero and standard deviation of 1.[25] In the equation generating the teacher effect τ_{jt} (Equation K2), the coefficient on the unobserved term u' is chosen so that the teacher effect has a standard deviation of 0.16, following the observed data from District D in Year 7. After all predictor variables and unobserved terms were randomly generated and constrained to having the characteristics described in Equations K3 through K5 (and independent of each other unless otherwise noted (as in Equation K5)), Equations K1 and K2 were combined to create a data generation function for current achievement at the student level. Again approximating the data from District D in Year

[25] While the free- and reduced lunch variable in most administrative data-sets is binary or ordinal, for the ease of simulating data, I employ a continuous, normally distributed variable in its place. However, the findings here should broadly generalize to scenarios with other binary or ordinal covariates.

7, the data was constructed such that 8,100 student observations were nested in 270 classes nested in 45 teacher – each teacher assigned to 6 classes, each consisting of 30 students.

Whereas "true" values of student prior achievement are used in the data generation Equation K1, "observed" values of student prior achievement are constructed as in Equation K6 such that the observed variables maintain a mean of zero and standard deviation of one, as in the real data used in this dissertation. The $achievement_{error}$ variable was constructed so as to be uncorrelated with all other exogenous variables. The $achievment_{observed}$ variable is constructed as to have a mean of zerp and standard deviation of 1.

$$Achievement_{observed} = (\sqrt{Reliability_{Achievement}})\, Achievement_{true_{jt}} + (\sqrt{1 - Reliability_{Achievement}})\, Achievement_{error_{jt}} \quad [K6]$$

After the $achievement_{observed}$ variable is constructed for each observation, a class-level average achievement is constructed ($\overline{Achievement_{observed}}_{jkt-1}$), and both of these variables are used as regressors predicting student-level achievement (Equation K7). The resulting residual (ε_{it}) is then averaged by teacher to create the estimated teacher-effect or teacher value-added, $\hat{\tau}_{jt}$ (Equation K8).

$$Achievement_{it} = \beta_0 + \beta_1 FRL_{it} + \beta_2 \overline{FRL}_{jkt} + \beta_3 Achievement_{observed_{it-1}} + \beta_4 \overline{Achievement_{observed}}_{jkt-1} + \varepsilon_{it} \quad [K7]$$

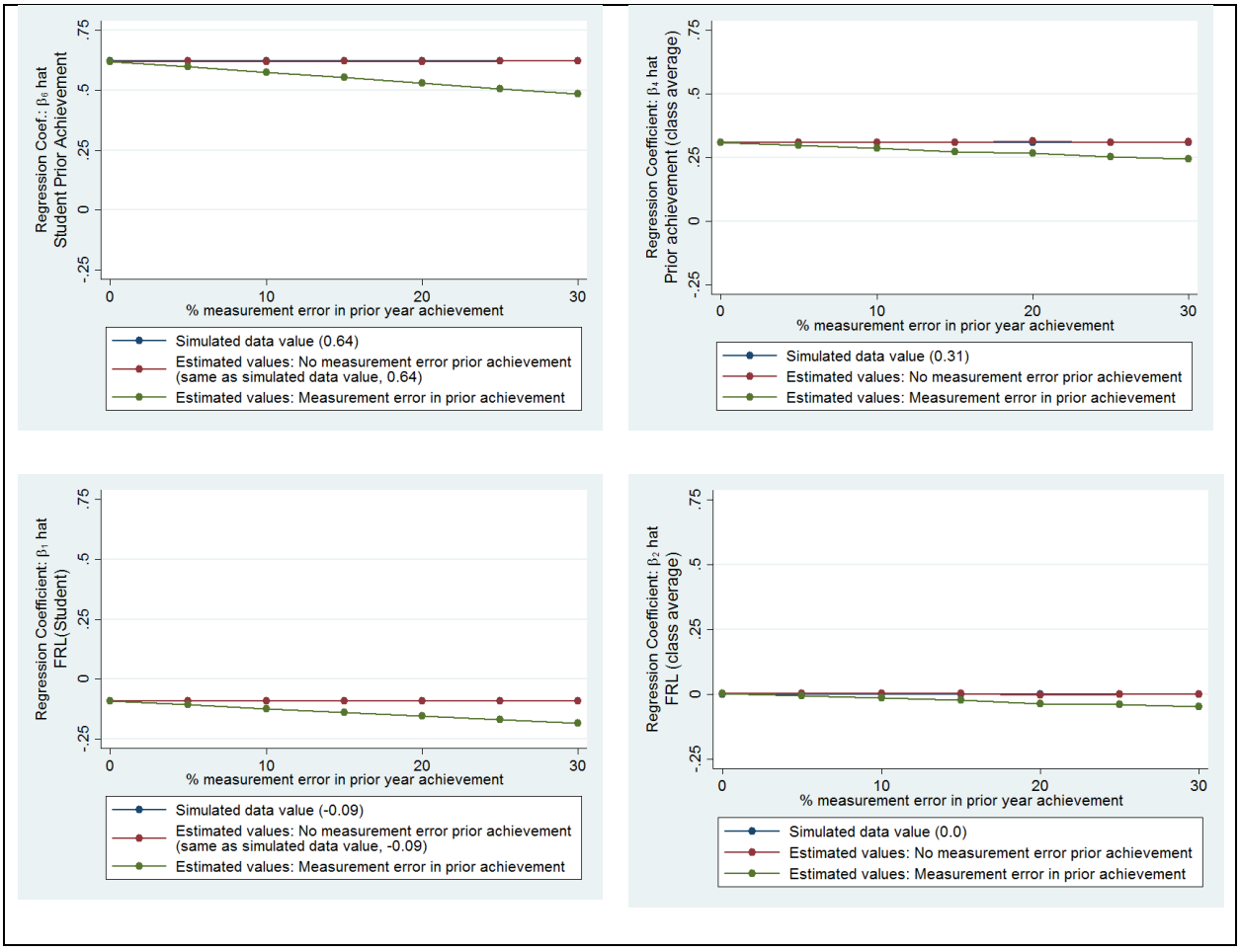
$$\hat{\tau}_{jt} = \bar{\varepsilon}_{it} \quad [K8]$$

For each of the seven values for reliability of student prior achievement tested by this simulation, 1,000 data sets following the parameters described in Equations K1 through K5 were randomly generated, resulting in 1,000 regressions and 1,000 estimates of the regression coefficients in Equation K7 and 1,000 estimates of teacher value-added $\hat{\tau}_{jt}$ for each of the seven levels of student prior achievement reliability (with reliability ranging from 70% to 100% by increments of 5%). Each of these coefficients and teacher value-added estimates were averaged within reliability level and reported below.

Student Prior Achievement Reliability: Results

Several changes are noted in the regression coefficients as measurement error is added into student prior level achievement. The coefficient on student prior engagement (β_4) becomes less positive as more measurement error is introduced into the variable, as is expected given that measurement error in dependent variables attenuates estimates of the relationship towards zero (Isenberg & Hock, 2010; Wooldridge, 2005). Similarly, the coefficient of the same variable aggregated and averaged at the classroom level – which has a positive coefficient in the production function for current-year student achievement (Equation K1) – also demonstrates attenuation towards zero as the percentage of measurement error in the aggregated variable is increased. The student-level FRL variable is negatively correlated with both prior student achievement and current-year student achievement. Because prior student achievement is positively correlated with current year achievement, when the two previously mentioned negative correlations are taken into account, it can be predicted that omitting the prior student achievement variable from the regression equation would introduce omitted variable bias causing the negative coefficient on the FRL variable to be biased downward (Wooldridge, 2005). This is consistent with the interpretation that controlling for aggregate prior achievement at the class level may serve as a form of correction for measurement error in the prior test score. If we conceive of increasing the proportion of measurement error in the FRL variable as omitting the variable by degree, then it becomes clear that the more measurement error that is introduced in the prior achievement variable, then the more negative the already negative coefficient on the FRL variable becomes.

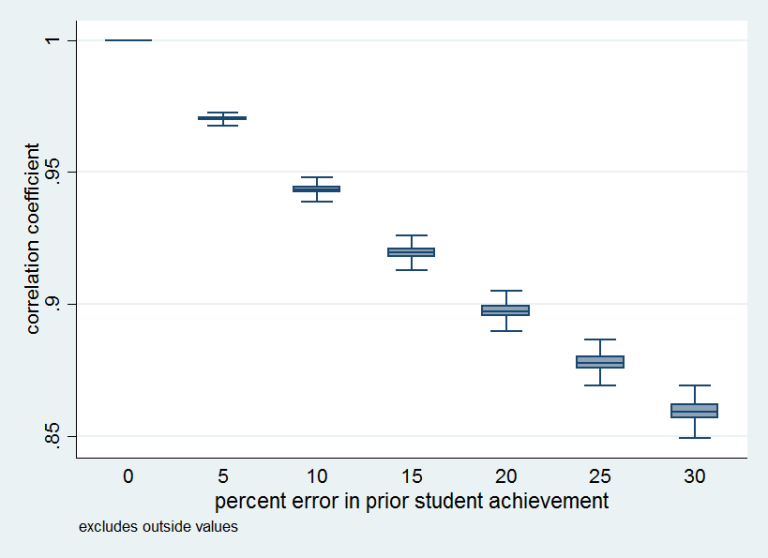
Figure K1a-d. Changes in estimated regression coefficients of predictor variables as percent measurement error increases. Each point estimate represents the averaged coefficient over 1,000 replications using simulated data. Graphs displayed are for the following predictor variables: (a) student prior achievement, (b) student prior achievement (class average), (c) student FRL, and (d) FRL (class average). In each of these graphs, at the scale chosen here for comparison, the lines plotting the value used to simulate the data are indistinguishable from the estimated values given no measurement error in student prior achievement.



However, for this regression predicting current year student achievement, none of the estimated regression coefficients are of primary interest. Instead, of primary interest are the residuals which result are recovered and then averaged within teacher to estimate teacher contribution to student learning. Consequently, what it of more interest is not the bias or change in the regression coefficients brought about by increasing measurement error, but instead the change in the residuals. The correlations between the unobserved student effect u_{it} and the calculated residuals ε_{it} are consistently 1.0 when no error is present in the predictor variables;

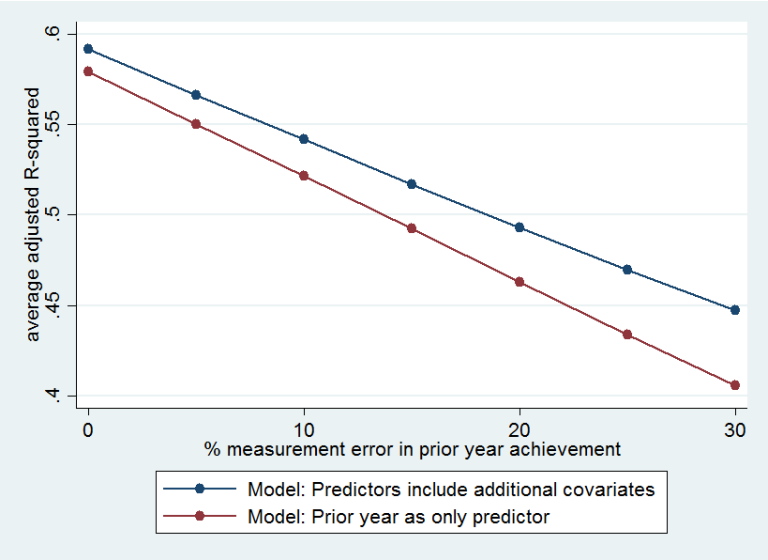
this correlation averages 0.859 across the 1,000 replications when reliability of prior student achievement is 0.70. The spread of these correlations across the 1,000 replications also increases as reliability decreases (see Figure K2).

Figure K2. Distribution of correlation of unobserved student effect u_{it} and the calculated residuals ε_{it} (n=1,000) at different levels of reliability of the prior student achievement variable.



The analysis here also suggests that including predictor covariates which are correlated with both prior- and current student achievement mitigates some of the reductions of predictiveness of the model (i.e., reductions of model fit) which is brought about by increasing percentages of error in the observed prior achievement variable. As is seen in Figure K3, compared to regressions with prior student achievement as the sole predictor variable, regressions which include additional predictor variables have better model fit but also result in model fit statistics which decrease more slowly as reliability in student prior achievement decreases.

Figure K3. Average adjusted R-squared for a regression model with prior year as the only predictor, compared to regression models with additional covariates as predictors, at different levels of reliability of student prior achievement. R-squared values plotted are averages of 1,000 replications on different simulated datasets.



Classroom Observation Reliability: Results

After the simulated student-level data was analyzed as described above, an analysis of the patterns of data associated with changing reliability of the classroom observation measure was conducted on teacher-level data. After the process described above, in which the estimated teacher effect $\hat{\tau}_{jt}$ is calculated, this variable becomes a dependent variable, which is predicted by the observed classroom variable, as in Equation K10:

$$\hat{\tau}_{jt} = \beta_5 + \beta_6 IQA_{observed_{jt}} + \varepsilon'_{jt} \qquad [K10]$$

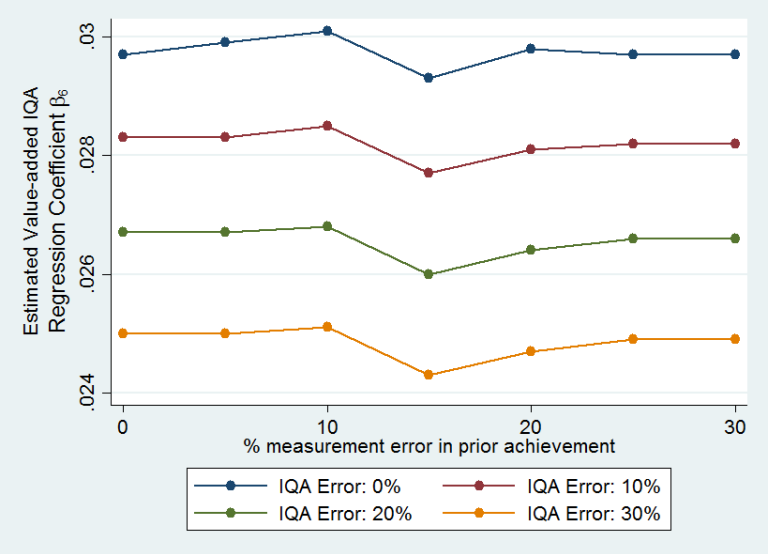
As with the observed prior student achievement variable, the IQA observed variable is constructed by introducing noise into a "true" variable used for data generation, with the amount of noise varying from zero to 30 percent of the total variance (i.e. reliability ranging from 100 percent to 70 percent).

$$IQA_{observed} = \left(\sqrt{Reliability_{IQA}}\right) IQA_{true\,jt} + \left(\sqrt{1 - Reliability_{IQA}}\right) IQA_{error\,jt} \quad [K11]$$

For each of the seven levels of reliability of prior student achievement investigated, there were 1,000 replications, each consisting of 45 teacher observations. In order to describe any possible interactions in student test score reliability and teacher observation reliability on the estimates of β_6, the regression coefficient describing the relationship between classroom observation and estimated teacher value-added, the range of teacher observation reliabilities was tested within each of the values of student achievement reliability.

Results from these regressions reveal that as more measurement error is introduced into the classroom observation measure, the weaker the relationship with estimated teacher effectiveness. As reliability in the IQA observational measure moves from 1.0 to 0.7, the average estimate of the coefficient on the IQA variable decreases from 0.297 to 0.250, a decrease in effect size of approximately 16 percent. By comparison, the influence of the reliability of the student prior achievement variable on estimates of the IQA coefficient appear to be unsystematic and relatively small compared to the influence of the reliability of the IQA measurement (see Figure K4).

Figure K4. Changes in estimates of the value-added – IQA coefficient by reliability in the student prior achievement variable, by reliability of the IQA observed variable.



REFERENCES

Aaronson, D., Barrow, L. and Sanders, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25,* 95–135.

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14.

Alliance for Excellent Education (2014, February). *New assessments: A guide for state policymakers*. Washington, DC: Alliance for Excellent Education. Retrieved from https://all4ed.org/wp-content/uploads/2014/02/AssessmentsStatePolicyGuide.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME]. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Appelgate, M. & Rosenquist, B. (2016). *As Common Core takes hold: Changes in teachers' mathematics curriculum use.* Presented at the National Council of Teachers of Mathematics national conference in San Francisco, California.

Artiles, A. J., Harry, B., Reschly, D. J., & Chinn, P. C. (2002). Over-identification of students of color in special education: A critical overview. *Multicultural Perspectives, 4*(1), 3-10.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Baker, E. L. (2004). *CSE Report 645: Aligning curriculum, standards, and Assessments: Fulfilling the promise of school reform.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers.* EPI Briefing Paper# 278. Economic Policy Institute.

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are NCLB's measures, incentives, and improvement strategies the right ones for the Nation's low-performing high schools?. *American Educational Research Journal, 44*(3), 559-593.

Ball, D. L. (1995). *Developing Mathematics Reform: What Don't We Know about Teacher Learning--But Would Make Good Working Hypotheses?* NCRTL Craft Paper 95-4.

Ball, D. L., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *The Elementary School Journal, 109*(5), 458-474.

Ballou, D. (2002). Sizing up test scores. *Education next, 2*(2).

Ballou, D. (2005). Value-added assessment: lessons from Tennessee. In R. LISSITZ (ed.), *Value-added Models in Education: Theory and Application.* Maple Grove, MN: JAM Press.

Ballou, D., & Springer, M. G. (2015). Using Student Test Scores to Measure Teacher Performance Some Problems in the Design and Implementation of Evaluation Systems. *Educational Researcher, 44*(2), 77-86.

Ballou, D., Mokher, C. G., & Cavalluzzo, L. (2012*). Using value-added assessment for personnel decisions: How omitted variables and model specification influence teachers' outcomes.* Unpublished manuscript, Department of Leadership, Policy, and Organizations, Vanderbilt University, Nashville, Tennessee.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics, 29*(1), 37-65.

Banilower, E. R., Boyd, S. E., Pasley, J. D., & Weiss, I. R. (2006). *Lessons from a Decade of Mathematics and Science Reform: A Capstone Report for the Local Systemic Change through Teacher Enhancement Initiative.* Chapel Hill: Horizon Research, Inc.

Battista, M. T. (1999). The mathematical miseducation of America's youth. *Phi Delta Kappan, 80*(6), 424.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571). John Wiley & Sons.

Betebenner, D. (2007). *Estimation of Student Growth Percentiles for the Colorado Student Assessment Program*. Dover, NH: National Center for the Improvement of Educational Assessment.

Bidwell, A. (2014, August 20). Common Core support if free fall. *US World & News Reports.* Retrieved from the USW&NR website: http://www.usnews.com/news/articles/2014/08/20/common-core-support-waning-most-now-oppose-standards-national-surveys-show

Boaler, J. (2002). *Experiencing school mathematics: Traditional and reform approaches to teaching and their impact on student learning.* New York: Routledge.

Boardman, A.E., Murnane, R.J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education 52*(2),113–121.

Bollen, K. A., & Jackman, R. W. (1990*). Regression diagnostics: An expository treatment of outliers and influential cases*. In J. Fox & J.S. Long Modern methods of data analysis, 257-291.

Booher-Jennings, J. (2005). Below the bubble:"Educational triage" and the Texas accountability system. *American educational research journal, 42*(2), 231-268.

Booker, L.N., (2014). *Examining the Development of Beginning Middle School Math Teachers' Practices and their Relationship with the Teachers' Effectiveness*. Unpublished doctoral thesis. Nashville, Tennessee: Vanderbilt University.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2009). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*, 417-436.

Borman, K. M. (2005). *Meaningful urban education reform: Confronting the learning crisis in mathematics and science*. Albany, NY: SUNY Press

Boston, M. (2012). Assessing Instructional Quality in Mathematics. *The Elementary School Journal, 113*(1), 76-104.

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of the instructional quality assessment toolkit CSE Technical Report* (Vol. 672). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Boyd, D., Lankford, R. H., Loeb, S., & Wycoff, J. (2003). Understanding teacher labor markets: Implications for educational equity. In M. L. Plecki & D. H. Monk (Eds.), *School finance and teacher quality: American Education Finance Association 2003 yearbook* (pp. 55–84). Larchmont, NY: Eye on Education.

Boyd, D., Lankford, R. H., Loeb, S., & Wycoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics 38*(6), 629-663.

Braun, H. I. (2005). *Policy Information Perspective: Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.*. Princeton, NJ Educational Testing Service.

Brennan, R.L. (2001). *Generalizability theory.* New York: Springer-Verlag

Bryk, Anthony S., and Barbara Schneider. 2002. *Trust in Schools. A Core Resource for Improvement.* New York: Russell Sage Foundation.

Buil, J. M., Koot, H. M., Olthof, T., Nelson, K. A., & Lier, P. A. (2015). DRD4 genotype and the developmental link of peer social preference with conduct problems and prosocial behavior across ages 9–12 years. *Journal of youth and adolescence, 44*(7), 1360-1378.

California Department of Education. (1992). *Mathematics Framework for California Public Schools.* Sacramento, CA: Author

Campbell, D. T. (2010). Assessing the impact of planned social change. *Journal of MultiDisciplinary Evaluation, 7*(15), 3-43.

Cavalluzzo, L. C. (2004). *Is National Board Certification an effective signal of teacher quality?* (Report No. IPR 11204). Alexandria, VA: CNA Corporation. Retrieved from the ERIC website: http://files.eric.ed.gov/fulltext/ED485515.pdf

Carlos, L., & Kirst, M. (1997). *California Curriculum Policy in the 1990's:" We Don't Have To Be in Front To Lead."* Paper presented at the Annual Meeting of the American Educational Research Association in Chicago, IL.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. L. Cheng & A. Curtis (eds.) *Washback in language testing: Research contexts and methods*, (pp. 3-17). New York: Routledge

Cheong, Y. F., Fotiu, R. P., & Raudenbush, S. W. (2001). Efficiency and robustness of alternative estimators for two-and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics, 26*(4), 411-429.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.* NBER Working Paper No. 17699.

Chingos, M. M. (2013). *Standardized testing and the common core standards: You get what you pay for?*. Washington: Brookings Institution.

Chingos, M.M. (2015, February 2). *Up Front: Testing costs a drop in the bucket [*Blog post]. Retrieved from https://www.brookings.edu/blog/up-front/2015/02/02/testing-costs-a-drop-in-the-bucket/

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review, 30*(3), 449-465.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs (NCEE 2013-4015).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Clotfelter, C., Ladd, H. F., & Vigdor, J. (2004). *Teacher quality and minority achievement gaps.* Durham, NC: Terry Sanford Institute of Public Policy.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. Journal of Human Resources, 41(4), 778-820.

Cobb, P., & Jackson, K. (2011). Towards an Empirically Grounded Theory of Action for Improving the Quality of Mathematics Teaching at Scale. *Mathematics Teacher Education and Development, 13*(1), 6-33.

Cobb, P., & Smith, T. (2008). District development as a means of improving mathematics teaching and learning at scale. *International handbook of mathematics teacher education, 3,* 231-254.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3-12.

Cogan, L. S., Schmidt, W. H., & Wiley, D. E. (2001). Who takes what math and in which track? Using TIMSS to characterize US students' eighth-grade mathematics learning opportunities. *Educational Evaluation and Policy Analysis, 23*(4), 323-341.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, D., & Hill, H. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. Philadelphia: Consortium for Policy Research in Education

Cohen-Vogel, L. (2011). "Staffing to the test" are today's school personnel practices evidence based?. *Educational Evaluation and Policy Analysis, 33*(4), 483-505.

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS education value-added assessment system (EVAAS®). *Education policy analysis archives, 22*, 98.

Common Core State Standards Initiative (n.d.). *Development process.* Retrieved from the Common Core State Standards Initiative website: http://www.corestandards.org/about-the-standards/development-process/

Conley , D.T. (2012). A complete definition of college and career readiness. Eugene, OR: Educational Policy Improvement Center

Conley, D. T. (2014). *The Common Core State Standards: Insight into their development and purpose*. Washington, DC: Council of Chief State School Officers.

Conley, D. T., Drummond, K. V., de Gonzalez, A., Seburn, M., Stout, O., & Rooseboom, J. (2011). *Lining up: The Relationship between the Common Core State Standards and Five Sets of Comparison Standards.* Eugene, OR: Educational Policy Improvement Center (NJ1).

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy 8*(3), 418-434.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). *Teacher Effectiveness on High-and Low-Stakes Tests.* Presented at the 2011 conference for the Society for Research on Educational Effectiveness. Retrieved from https://www.sree.org/conferences/2011/program/downloads/abstracts/131.pdf

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation, 10*(7), 1-9.

Cowan, J., & Goldhaber, D. (2016). National Board certification and teacher effectiveness: evidence from Washington State. *Journal of Research on Educational Effectiveness, 9*(3), 233-258.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137-163.

Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system* (NBER Working Paper #12286). Cambridge, MA: National Bureau for Economic Research.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Measurement, 12*, 1–22.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* Dallas, TX: National Staff Development Council.

Doyle, W. R., Lee, J., & Nguyen, T. D. (2017). *Impact of Need-Based Financial Aid on Student Persistence at Public Institutions: An Application of the Regression-Discontinuity Design.* Available at SSRN: https://ssrn.com/abstract=2930286

Duke, D. L. (2012). Tinkering and turnarounds: Understanding the contemporary campaign to improve low-performing schools. *Journal of Education for Students Placed at Risk 17*(1-2), 9-24.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education,* 4 (4), 289-303.

Dynarski, M. (2016, December 8*). Report: Teacher observations have been a waste of time and money.* Washington, DC: Brookings Institution

Elliott, R., Kazemi, E., Lesseig, K., Mumme, J., Carroll, C., & Kelley-Petersen, M. (2009). Conceptualizing the work of leading mathematical tasks in professional development. *Journal of Teacher Education, 60*, 364-379.

Elliott, S. W., & Hout, M. (Eds.). (2011). *Incentives and test-based accountability in education.* Washington, DC: National Academies Press.

Elmore, R., & Fuhrman, S. (1995). Opportunity-to-learn standards and the state role in education. *The Teachers College Record, 96*(3), 432-457.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement.* 3rd ed., pp. 105–146). New York, NY: American Council on Education.

Ferguson, R. F., & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265–298). Washington, DC: Brookings.

Figlio, D., & Getzler, L. (2002). *Accountability, ability and disability: Gaming the system?* (National Bureau for Economic Research Working Paper 9307). Cambridge, MA: National Bureau for Economic Research.

Finn, B. (2015). *Measuring Motivation in Low-Stakes Assessments.* ETS Research Report Series, 2015(2), 1-17.

Finn, C.E., Julian, L. & Petrilli, M.J. (2006, August). 2006 *The State of State Standards.* Washington, DC: The Thomas B. Fordham Foundation. Retrieved from the Thomas B. Fordham website: http://edexcellence.net/publications/soss2006.html6

Finn Jr, C. E., Petrilli, M. J., & Vanourek, G. (1998). *The State of State Standards.* Washington, DC: The Thomas B. Fordham Foundation

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and teacher education, 19*(6), 643-658.

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap?. *The High School Journal*, *91*(1), 29-42.

Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 225-256). Greenwich, CT: Information Age.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher-.18* (9), 27-32.

Fullan, M. (2000). The return of large-scale reform. *Journal of educational change, 1*(1), 5-27.

Fullan, M. (2009). Large-scale reform comes of age. *Journal of educational change, 10*(2-3), 101-113.

García, G. E., & Pearson, P. D. (1994). Assessment and diversity. *Review of research in education*, 337-391.

Garrison, A.L. (2013). *Understanding Teacher and contextual factors that influence the enactment of cognitively demanding mathematics tasks*. (Doctoral Dissertation). Retrieved from the Vanderbilt University Library website: http://etd.library.vanderbilt.edu/available/etd-06072013-080515/unrestricted/GarrisonDissertationFinal.pdf

Gates Foundation. (2010). *Learning from Teaching: Initial findings from the measures of effective teaching project.* Retrieved from the MET Project website: http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all?. *Language Testing, 26*(4), 507-531.

Goddard, Y. L., Goddard, R.D., & Tschannen- Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record, 109* (4): 877–896.

Goertz, M. E. & Duffy, M. C.. (2001*). Assessment and Accountability Systems in the 50 States: 1999-2000.* CPRE Research Reports. Retrieved from http://repository.upenn.edu/cpre_researchreports/13

Goertz, M. E., Floden, R. E., & O'Day, J. A. (1996). *Studies of education reform: Systemic reform (Vol. 1).* Washington, DC: Dept. of Education.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1), 52-55.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics, 7*(3), 199-208.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129–145.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *The American Economic Review*, 250-255.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589-612.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28-39.

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

Greene, J. P. (2002). The business model. *Education Next, 2*(2). Retrieved from http://educationnext.org/files/ednext20022_20.pdf

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The Test Matters The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher,43*(6), 293-303.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted?. *Education Finance and Policy*.

Gutstein, E. (2003). Teaching and learning mathematics for social justice in an urban, Latino school. *Journal for Research in Mathematics Education*, 37-73.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.) *Educational Measurement* (pp. 65-110). Westport, CT: Prager

Hamilton, L.S. (2003). Assessment as a policy tool. *Review of research in education*, 25-68.

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. *Educational evaluation and policy analysis, 25*(1), 1-29.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education.* Santa Monica, CA: Rand Corporation.

Hamilton, L.S., Stecher, B., Marsh, J. A., Sloan McCombs, J,, Robyn, A., Russell, J, et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states.* Santa Monica, CA: RAND.

Hamilton, L.S., Stecher, B.M., & Yuan, K. (2008). *Standards-Based Reform in the United States: History, Research, and Future Directions.* Santa Monica, CA: RAND

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012). Standards-based accountability in the United States. *Education Inquiry, 3*(2), 149-170.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (Working Paper No. 11154). Cambridge, MA: National Bureau for Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*(2), 267-271.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798-812.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In M. Hersen (Ed.), *Comprehensive Handbook of Psychological Assessment* (Vol. 3, pp.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York: Routledge.

Havlicek, L. L., & Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. *Psychological Bulletin, 84*(2), 373.

Heck, D. J., Banilower, E. R., Weiss, I. R., & Rosenberg, S. L. (2008). Studying the effects of professional development: The case of the NSF's local systemic change through teacher enhancement initiative. *Journal for Research in Mathematics Education*, 113-152.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. Journal for *Research in Mathematics Education, 28*, 524-549.

Henry, G., Rose, R., & Lauen, D. (2014). Are value-added models good enough for teacher evaluations? Assessing commonly used models with simulated and actual data. *Investigaciones de Economía de la Educación* volume 9, 9, 383-405.

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability* (pp 141-166). New York: Teachers College Press.

Herman, J.L. (2008). Accountability and assessment: Is public interest in K-12 education being served. In K. E. Ryan and L. A. Shepard (Eds.) *The future of test-based educational accountability.* New York: Taylor & Francis

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, *3*(1), 1-10.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., ... & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis, 27*(2), 111-132.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences.* Independence, KY: Wadsworth/Cengage

Holcombe, R., Jennings, J., & Koretz, D.(2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation,* 163-189. Greenwich, CT: Information Age Publishing

Horn, I. S. (2006). Lessons learned from detracked mathematics departments. *Theory Into Practice, 45*(1), 72-81.

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal, 47*(1), 181-217.

Isenberg, E., & Hock, H. (2010). *Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools. Final Report.* Mathematica Policy Research, Inc.

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics, 95*(4), 1096-1116.

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (No. w18624). National Bureau of Economic Research.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.

Jackson, K., Garrison, A., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education, 44*(4), 646-682.

Jacob, B. (2005). Testing, accountability, and incentives: The impact of high-stakes testing in Chicago public schools. *Journal of Public Economics, 89*(5/6), 761–796.

Jacob, B., & Levitt, S. (2005). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843–877.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of labor Economics, 26*(1), 101-136.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of human resources, 45*(4), 915-943.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the Test" in the NCLB Era How Test Predictability Affects Our Understanding of Student Performance. *Educational Researcher*, 0013189X14554449.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education, 83*(2), 135-159.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615-631.

Kane, T. J., & Staiger, D. O. (2008*). Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Brief. MET Project. *Bill & Melinda Gates Foundation*.

Kazemi, E., Franke, M., & Lampert, M. (2009). Developing pedagogies in teacher education to support novice teachers' ability to enact ambitious instruction. In *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 12-30).

Keiser, J. M., & Lambdin, D. V. (1996). The clock is ticking: Time constraint issues in mathematics teaching reform. *The Journal of Educational Research, 90*(1), 23-31.

Keith, T.Z. (2006). *Multiple regression and beyond.* Boston: Allyn and Bacon

Keith, T.Z. & Cool, V.A. (1992). Testing models of school learning: Effects of quality of instruction, motivation, academic coursework, and homework on academic achievement. *School Psychology Quarterly 7* (207–226).

Kilpatrick, J., Martin, W. G., & Schifter, D. (Eds.). (2003). *A research companion to principles and standards for school mathematics.* Reston, VA: National Council of Teachers of English

Kiplinger, V. L. (2008). Reliability of large scale assessment and accountability systems. In K. E. Ryan and L. A. Shepard (Eds.) *The future of test-based educational accountability* (pp. 93-114). New York: Taylor & Francis

Klein, S. P., Hamilton, L., McCaffrey, D. F., & Stecher, B. (2000). What do test scores in Texas tell us?. *Education policy analysis archives, 8*, 49.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Nashville, TN: National Center on Performance Incentives, Vanderbilt, Peabody College.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *The BE Journal of Economic Analysis & Policy, 12*(1).

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. *Yearbook of the National Society for the Study of Education, 104*(2), 99-118.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us.* Boston: Harvard University Press.

Koretz, D., S. Barron, K. Mitchell, and B. Stecher. (1996). *The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS).* MR-792-PCT/FF. Santa Monica, Calif.: RAND.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational evaluation and policy analysis, 25*(3), 287-298.

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy, 12*(2), 241-279.

Lampert, M., Beasley, H., Ghousseini, H., Kazemi, E., & Franke, M. L. (2010). Using designed instructional activities to enable novices to manage ambitious mathematics teaching. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 129-141). New York: Springer.

Lampert, M., & Graziani, F. (2009). Instructional activities as a tool for teachers' and teacher educators' learning in and for practice. *Elementary School Journal, 109*(5), 491-509.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

Le, V. N., Lockwood, J. R., Stecher, B. M., Hamilton, L. S., & Martinez, J. F. (2009). A longitudinal investigation of the relationship between teachers' self-reports of reform-oriented

instruction and mathematics and science achievement. *Educational Evaluation and Policy Analysis, 31*(3), 200-220.

Le, V. N., Stecher, B. M., Lockwood, J. R., Hamilton, L. S., & Robyn, A. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement.* :Santa Monica, CA: RAND Corporation.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher, 43*(5), 242-252.

Leong, Y. H., & Chick, H. L. (2011). Time pressure and instructional choices when teaching mathematics. Mathematics *Education Research Journal, 23*(3), 347.

Linn, R. L. (2000). Assessments and accountability. *Educational researcher, 29*(2), 4-16.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan and L. A. Shepard (Eds.) *The future of test-based educational accountability*. New York: Taylor & Francis

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527–537. doi:10.1016/j.learninstruc.2008.11.001.

Lipscomb, S., Gill, B., Booker, K., & Johnson, M. (2010). *Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results. Final Report*. Cambridge, MA: Mathematica Policy Research, Inc.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22-52.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement, 44*(1), 47-67.

Lockwood, J.R., McCaffrey, D.F., & Savage (2017, March 2). *Errors-in-Variables Regression: Why Stata's –eivreg- is wrong and what to do instead.* A paper presented at the Society for Research in Educational Effectiveness (SREE) conference in Washington, DC. Retrieved from the SREE website: https://www.sree.org/conferences/2017s/program/download.php?id=1910&part=1910_0.pdf&item=slide

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education, 123*(4), 615-646.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12* (SPA8954759). Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Manouchehri, A. (1998). Mathematics curriculum reform and teachers: What are the dilemmas?. *Journal of teacher education, 49*(4), 276-286.

Marsh, D. D., & Odden, A. R. (1991). Implementation of the California mathematics and science curriculum frameworks. In A.R. Odden (Ed.) *Education policy implementation*, Albany, NY: State University of New York Press, 219-240.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., & Peng, A. (2011*). A big apple for educators: New York City's experiment with schoolwide performance bonuses* (MG-1114-FPS). Santa Monica, CA: RAND Corporation Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361

Massell, D. (1994). Achieving consensus: Setting the agenda for state curriculum reform. In R.F. Elmore & S.H. Fuhrman (Eds.), *The governance of curriculum: 1994 yearbook of the Association for Supervision and Curriculum Development* (pp. 84-108). Alexandria, VA: Association for Supervision and Curriculum Development.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal, 108*(4), 293-312.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. B. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: a pilot study of the instructional quality assessment.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004).*Evaluating Value-Added Models for Teacher Accountability. Monograph*. Santa Monica, CA: RAND Corporation

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

McClain, K. (2002). Teacher's and students' understanding: The role of tool use in communication. *Journal of the Learning Sciences, 11*, 217-249.

McLaughlin, M. W., & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement (Vol. 45).* New York: Teachers College Press.

Massell, D. (2000). *The district role in building capacity: Four strategies. CPRE Policy Briefs, 1-7.* Philadelphia, PA: Consortium for Policy Research in Education

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports*. Office of Planning, Evaluation and Policy Development, US Department of Education. Retrieved from the ERIC website: http://files.eric.ed.gov/fulltext/ED516494.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Messick, S. (1990). Validity of test interpretation and use. *ETS Research Report Series, 1990*(1), 1487-1495.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125–145.

Morgen, S. (2017, February 7). *Bills would eliminate master's degree requirement for teachers.* Oak Grove, KY: The Eagle Post. Retrieved from http://www.kentuckynewera.com/ep/news/article_9e439ce4-edbc-11e6-a9bb-ff3c3e12ee58.html

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance,* 1-26.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills. Principles for educating children to thrive in a changing economy.* New York: The Free Press

National Center for Education Statistics. (2007). *Mapping 2005 state proficiency standards onto the NAEP scales (No. NCES 2007-482).* Washington, DC: U.S. Department of Education

National Commission on Excellence in Education (1983). *A nation at risk.* Washington, DC: The National Commission on Excellence in Education, US Department of Education.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Boston: Author.

National Council on Education Standards and Testing (1992*). Raising standards for American education.* Washington, DC: U.S. Government Printing Office.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000*). Principals and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2014). *Principles to Actions: Ensuring Mathematical Success for All*. Reston, VA: Author.

National Governors Association for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics.* Retrieved from the Common Core website: http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics.* Washington, D.C.: National Academies Press.

National Research Council. (2002). *Scientific research in education.* Washington, DC: National Academies Press.

National Research Council. (2004a). *Advancing Scientific Research in Education*. Washington, DC: The National Academies Press.

National Research Council. (2004b). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations.* Washington, DC: The National Academies Press.

National Research Council (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halper, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American psychologist, 51*(2), 77.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: new findings and theoretical developments. *American psychologist, 67*(2), 130.

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

O'Day, J.A & Smith, M.S. (1993). Systemic reform and educational opportunities. In S.H.Fuhrman (Ed.) *Designing coherent education policy: Improving the system.* San Francisco: Jossey-Bass

Olszyk, M. D., & Kessler, C. (2008). Emergency medicine in the VA: The battleship isturning. *Annals of Emergency Medicine 51*(5), 632-635.

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

Park, Y.S., Chen, J., Holtzman, S.L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R.C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New guidance from the Measures of Effective Teaching Project.* (pp. 383-410). San Francisco: Josey Bass

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3-14.

Polikoff, M. S. (2014). Does the test matter?: Evaluating teachers when tests differ in their sensitivity to instruction. In T. J. Kane, K. A. Kerr, & R.C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New guidance from the Measures of Effective Teaching Project.* (pp. 278-302). San Francisco: Josey Bass

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards?. *American Educational Research Journal, 48*(4), 965-995.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, 679-682.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational leadership, 56*, 8-16.

Popham, W.J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*, 146–155

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational researcher, 31*(7), 3-14.

Porter, A. C. (2006). Curriculum assessment. In G.L. Green, G. Camilli, P.B. Elmore (Eds.) *Handbook of complementary methods in education research* (pp. 141-159). New York: Routledge

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards the new US intended curriculum. *Educational Researcher, 40*(3), 103-116.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators.* Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*, 387–400.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding statistics: Statistical issues in psychology, education, and the social sciences, 2*(1), 13-43.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance & Policy*, *4*(4), 492-519.

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75,* 211-246.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmark and alignment of standards and testing. *Educational Assessment*, *9*(1&2), 1–27.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute

Rice, J. K. (2010). *The Impact of Teacher Experience: Examining the Evidence and Policy Implications. Brief No. 11.* Washington, DC: National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158-176.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2), 261-66..

Rosenquist, B.A., Henrick, E., & Smith, T.M. (2013, April). *Instructional leadership, teacher experience, and districts' supports for teachers: Teacher retention in three urban districts.* Presented at the American Educational Research Association (AERA) annual conference in San Francisco, CA.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools. *Teachers College Record, 104*(8), 1525–1567.

Rubin, Donald B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association 81* (396): 961–62.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691-712.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Ruzek, E. A., Hafen, C. A., Hamre, B. K., & Pianta, R. C. (2014). Combining classroom observations and value added for the evaluation and professional development of teachers. In T. J. Kane, K. A. Kerr, & R.C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New guidance from the Measures of Effective Teaching Project.* (pp. 205-233). San Francisco: Josey Bass

SAS Institute, Inc. (2015) *Solution overview: SAS® EVAAS ® for K-12.* Retreived from the SAS website: http://www.sas.com/content/dam/SAS/en_us/doc/overviewbrochure/sas-evaas-k12-104570.pdf

Sampson, T. (2010, February 11). *New rating system to begin in 2012.* Washington, DC: McClatchy-Tribune Business News

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review, 38*, 9-23.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction, 17*, 1-24.

SCANS Commission. (1991). *What work requires of schools: A SCANS Report for America 2000.* Washington, DC: The Secretary's Commission on Achieving Necessary Skills, US Department of Labor

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM, 48*(1-2), 29-40.

Schmidt, R. A. (2013). *Unpacking tracking: The role of instruction, teacher beliefs and supplemental courses in the relationship between tracking and student achievement* (Doctoral dissertation, Vanderbilt University).

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *Splintered vision: An investigation of US mathematics and science education.* Norwel, MA: Kluwer Academic Publishers.

Schoenfeld, A. H. (2007). Issues and Tensions in the Assessment of Mathematical Proficiency. In A. H.Schoenfeld (eds.) *Assessing mathematical proficiency*: *Mathematical Sciences Research Institute Publications (Book 53).* Cambridge, UK: Cambridge University Press.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education, 49*, 413-430.

Sharpe, C.J.D., Rosenquist, B.A., & Kern, E. (n.d.) *Unpublished working paper: Marshaling Evidence for Increased Test Rigor for Career and College Ready Standards-aligned State Assessments*

Shepard, L.A. (2008). A brief history of accountability testing, 1965-2007. In K. E. Ryan and L. A. Shepard (Eds.) *The future of test-based educational accountability*. New York: Taylor & Francis

Smith, M.S. & O'Day, J.A. (1991). Systemic School Reform. In S. Fuhrman and B. Malen (eds.), *The Politics of Curriculum and Testing.* Bristol, PA: Falmer Press

Spillane, J. P., & Thompson, C. L. (1997). Reconstructing conceptions of local capacity: The local education agency's capacity for ambitious instructional reform. *Educational Evaluation and Policy Analysis*, *19*(2), 185-203.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1-27.

Springer, M. G., Pane, J. F., Le, V. N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367-390.

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica, CA: RAND.

Stecher, B. (2010). *Performance Assessment in an Era of Standards-Based Educational Accountability.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classroom* (CRESST Report 525). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M. & Borko, H. (2002). Combining surveys and case studies to examine standards-based *educational reform* (CRESST Report 565). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stein, M. K. (2004). Studying the influence and impact of standards: The role of districts in teacher capacity. In J. Ferrini-Mundy & F. K. Lester (Eds.), *Proceedings of the National Council of Teachers of Mathematics Research Catalyst Conference*. Reston, VA: National Council of Teachers of Mathematics.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50-80.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development.* New York: Teachers College Press.

Stigler, J. W., & Hiebert, J. (1998). Teaching Is a Cultural Activity. *American Educator, 22*(4), 4-11.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87-100

Strutchens, M. E., Lubienski, S. T., McGraw, R., & Westbrook, S. K. (2000). NAEP findings regarding race and ethnicity: Students' performance, school experiences, attitudes and beliefs, and family influences. In P. Kloosterman & F.K. Lester, Jr. (eds). *Results and interpretations of the 1990 through 2000 mathematics assessments of the National Assessment of Educational Progress*. (pp. 269-304) Reston, VA: National Council of Teachers of Mathematics

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology, 90*(2), 269.

Tarr, J. E., Reys, R. E., Reys, B. J., Chavez, O., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. Journal for Research in Mathematics Education, 247-280.Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education.* Washington, DC: Council of Chief State School Officers

Texas Education Agency (N.D.). *STAAR media toolkit.* Retrieved from the Texas Education Agency (TEA) website: http://tea.texas.gov/index2.aspx?id=2147504081

Topol, B., Olson, J., & Roeber, E. (2012). *Getting to Higher-Quality Assessments: Evaluating Costs, Benefits and Investment Strategies.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education

The New Teacher Project (2015). *The Mirage: Confronting the hard truth about our quest for teacher development.* Brooklyn, NY: Author. Retrieved from The New Teacher Project website: http://tntp.org/assets/documents/TNTP-Mirage_2015.pdf

Thompson, D. R., & Senk, S. L. (2001). The effects of curriculum on achievement in second-year algebra: The example of the University of Chicago School Mathematics Project. *Journal for Research in Mathematics Education*, 58-84.

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., ... & Goble, L. (2015). *Understanding the Effect of KIPP as It Scales: Volume I, Impacts on Achievement and Other Outcomes. Final Report of KIPP's Investing in Innovation Grant Evaluation.* Washington, DC: Mathematica Policy Research, Inc

Ujifusa, A. (2017, September 18). Map: Tracking the Common Core State Standards. *Education Week.* Retrieved from the Education Week website: https://www.edweek.org/ew/section/multimedia/map-states-academic-standards-common-core-or.html

U.S. Department of Education (2009). *Standards and assessment peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author

Value-added Research Center (2015). *Projects*. Downloaded from the Value-Added Research Center website: http://varc.wceruw.org/Projects/

van Herpen, M., Van Praag, M., & Cools, K. (2005). The effects of performance measurement and compensation on motivation: An empirical study. *De Economist, 153*(3), 303-329.

Walkington, C. & Marder, M. (2014). Classroom observation and value-added models giver complementary information about quality of mathematics teaching. In T. J. Kane, K. A. Kerr, & R.C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New guidance from the Measures of Effective Teaching Project.* (pp. 234-277). San Francisco: Josey Bass

Walsh, E. & Isenberg, E. (2015). How does value added compare to student growth percentiles?, *Statistics and Public Policy, 2*(1)

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states (Research Monograph No. 18).* Madison, WI: National Institute for Science Education.

Weinstein, L. A. (1993). *Moving a battleship with your bare hands: Governing a university system.* Madison, WI: Magna Publications

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom*. Chapel Hill, NC: Horizon Research Inc.

Wilhelm, A. G., & Kim, S. (2015). Generalizing From Observations of Mathematics Teachers' Instructional Practice Using the Instructional Quality Assessment. *Journal for Research in Mathematics Education, 46*(3), 270-279.

Wiliam, D. (2007). *Sensitivity to instruction: the missing ingredient in largescale assessment systems?* Paper presented at the annual meeting of the International Association for Educational Assessment: Baku, Azerbaijan, September 2007

Wixson, K.K., Dutro, E., & Athan, R.G. (2003). The challenge of developing content standards. *Review of Research in Education, 27,* 69-107.

Wooldridge, J. M. (2005). *Introductory Econometrics: A Modern Approach.* South-Western College: Nashville, TN

Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. Santa Monica, CA: RAND Corporation. Retrieved from the RAND website: http://www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR967.pdf

Zohoori, N., & Savitz, D. A. (1997). Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Annals of epidemiology 7*(4), 251-257.