Are Practice-Based Teacher Evaluations and Teacher Effectiveness Linked in TNTP's *Performance Assessment System*?

APRIL 2012

Center for Education Policy Research
HARVARD UNIVERSITY

Center for Education Policy Research

The rapid accumulation of student achievement data by states and school districts represents an untapped national resource that promises to break longstanding stalemates in the American education policy debate. By partnering with education and policy leaders to help address their most pressing challenges, The Center for Education Policy Research (CEPR) represents a new paradigm in education research and has become a focal point for education policy researchers across Harvard and around the United States.

TNTP

TNTP (formerly The New Teacher Project) strives to end the injustice of educational inequality by providing excellent teachers to the students who need them most and by advancing policies and practices that ensure effective teaching in every classroom. A national nonprofit organization founded by teachers, TNTP is driven by the knowledge that effective teachers have a greater impact on student achievement than any other school factor. In response, TNTP develops customized programs and policy interventions that enable education leaders to find, develop and keep great teachers. Since its inception in 1997, TNTP has recruited or trained approximately 49,000 teachers - mainly through its highly selective Teaching Fellows programs - benefiting an estimated 8 million students. TNTP has also released a series of acclaimed studies of the policies and practices that affect the quality of the nation's teacher workforce, including The Widget Effect (2009) and Teacher Evaluation 2.0 (2010). Today TNTP is active in more than 25 cities, including 10 of the nation's 15 largest.

Acknowledgements

This study was funded by the Bill & Melinda Gates Foundation. The authors thank Dr. George Noell and Laura Boudreaux Pitre from the Louisiana Department of Education for their insight on project design and assistance compiling data, respectively. Ila Deshmukh Towery and Fiona Lin from TNTP provided their guidance and feedback on the components of PAS. Tim Daly and the TNTP team generously supplied their time and support for the project.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

Are Practice-Based Teacher Evaluations and Teacher Effectiveness Linked in TNTP's *Performance Assessment System*?

APRIL 2012

John H. Tyler, Brown University

Brian A. Jacob, University of Michigan

Shaun M. Dougherty, Harvard University

Havala J. Hanson, Harvard University

Jon B. Fullerton, Harvard University

Corinne M. Herlihy, Harvard University

Are Practice-Based Teacher Evaluations and Teacher Effectiveness Linked in TNTP's *Performance Assessment System*?

Recent research on the importance of teachers in promoting academic achievement has helped fuel an increased interest in and emphasis on teacher evaluation.[1] Based on evidence that has shown the importance of teachers in the education production process, the wide variation in the ability of different teachers to promote achievement, and the difficulty in identifying the most and least effective teachers, federal and state governmental agencies are encouraging school districts to develop new systems for evaluating teachers more rigorously than has traditionally been the case.[2] While such systems might incorporate student test scores in some manner, fewer than half of the nation's teachers could currently be evaluated solely on the basis of student test scores because of data limitations.[3] Such limitations suggest a central role for practice-based evaluation systems that rate teachers on the basis of their observed classroom practice and the artifacts of their teaching. A central question for any practice-based evaluation system, however, is to what extent are the valued elements of such a system related to student learning. Put another way, to what extent are teacher performance ratings on a practice-based evaluation system predictive of a teacher's ability to promote student achievement growth?

While several classroom observation rubrics have been validated against student achievement in research projects[4], to date, very few practice-based systems are able to answer this question in a practical setting with

---

[1] Evidence on the importance of teachers can be found in Aaronson, Borrow, and Sander (2007), Gordon, Kane, and Staiger (2006), Kane, Rockoff, and Staiger (2006), Rivkin, Hanushek, and Kain (2005), and Rockoff (2004).

[2] Weisberg, Daniel, Sexton, Susan, Mulhern, Jennifer, Keeling, David. THE WIDGET EFFECT: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness (TNTP, 2009). Also, see program description for Race to the Top, http://www2.ed.gov/programs/racetothetop/index.html.

[3] Using student test score growth to evaluate teachers requires the regular, annual testing of students. Except for rare exceptions, this type of testing regime currently exists only for students in grades 3-8 and only in math and reading.

[4] For information on classroom observation rubrics, see Kane, Thomas and Doug Staiger. Gathering Feedback for Teaching (Bill and Melinda Gates Foundation, 2012). For an example of classroom observation used for professional development, see Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011) "An interaction-based approach to enhancing secondary school instruction and student achievement." Science 333(6045):1034–37.

real consequences for teachers. A recent exception is a study by Kane, Taylor, Tyler, and Wooten (2011) demonstrating that teachers' scores on Cincinnati's practice-based Teacher Evaluation System are highly predictive of a teachers' ability to promote achievement growth.[5] This paper engages in a similar validation exercise for TNTP's Performance Assessment System (PAS), an evaluation tool developed to assess a prospective teacher's ability to provide effective instruction completed during their first year of teaching to determine initial certification. While there are similarities between Cincinnati's system and PAS, investigating PAS specifically is critical for three reasons. First, studies like Kane et al. are rare; there remains much marginal benefit in studying additional systems and settings to understand whether Cincinnati's results are common or uncommon. Second, while PAS evaluators determine a teacher's scores, they draw on a broader set of evidence than do Cincinnati's evaluators (e.g., parent and student surveys, student work). Finally, TNTP is a national organization with reach into many districts. As a result, there is the potential that the PAS could be quickly implemented and scaled across geographies. Given that the components of the PAS teacher evaluation system are characteristic of early proposals for the kinds of systems districts and states are being encouraged to adopt by federal grant programs (e.g., Race to the Top), a validation study of this system is both timely and relevant. And third, recent work by Taylor and Tyler (2012) finds that teachers who undergo evaluation in the Cincinnati system that has been validated against teacher value-added are more effective because of the evaluation.[6] Validating the PAS system is a first step in future work that examines the same question for PAS: does going through a PAS evaluation and certification process make teachers more effective?

Although our goal is to examine how well current methods of PAS-based evaluation identify effective teachers based on value-added, it is worth noting that TNTP may have other objectives in mind when using PAS to evaluate new teachers. With that reminder, this paper is focused on answering the following two questions:

- To what extent are PAS scores predictive of a teacher's ability to promote student achievement growth as measured by end-of-year state exams?
- Do some PAS framework areas and evidence sources provide more information about student outcomes than others?

[5] There is some earlier research on this topic by Milanowski (2004a, 2004b) and Holtzapple (2003) that also draws on data from Cincinnati.

[6] Value-added measures are estimated for teachers and represent average learning gains made by students who are all taught by the same teacher. For a more complete consideration of teacher value-added see, for example, Gordon, Kane, & Staiger, 2006; Sanders & Rivers, 1996.

I. TNTP and Louisiana's Practitioner Teacher Program

TNTP ADMINISTERS the Louisiana Practitioner Teacher Program (TNTP Academy - Louisiana), which is one of several non-traditional paths to teacher certification in Louisiana.[7] All TNTP Academy - Louisiana teachers participate in one of three programs that partner with TNTP: Teach Baton Rouge (TBR), which was started by TNTP in conjunction with the East Baton Rouge Parish (school district) ten years ago, but now is a separate entity; teachNOLA (tNOLA), which is also a subsidiary of TNTP; and Teach For America (TFA). Each of these three programs does their own recruiting, but all candidates from these programs enter the TNTP Academy - Louisiana certification pathway and are called Practitioner Teachers (PTs).

Earning a Level 1 certificate through the TNTP Academy - Louisiana program in Louisiana is designed to be a one-year process, and first requires that a PT qualify for a practitioner license, which enables them to teach while pursuing the more permanent certification. To earn a practitioner license, PTs must meet several criteria. First, PTs must be admitted to one of the three partner programs (generally in the spring or summer). Prior to commencing work with a partner program, they must also have a bachelor's degree with a minimum grade point average of 2.5 and must have passed all three of the exams that Louisiana requires for teacher certification. PTs must have no prior teaching license in Louisiana, nor are they allowed to have a BA or MA in education. In some cases candidates may have a certificate from another state, and in some instances they may also have teaching experience, though teachNOLA requires that applicants not be currently employed as teachers.[8]

Partner programs do not guarantee job placement and thus, once being admitted, all PTs must be hired by a school in their intended subject area, and begin teaching in the fall after being admitted to the partner program.

---

[7] In addition to the TNTP Academy - Louisiana program there are other alternative certification programs that exist in Louisiana, but they are not run by TNTP. For instance, the Louisiana Resource Center for Educators Practitioner Teacher Preparation program is among several listed in documentation made publically available by the state. Louisiana was the first state to implement a value added analysis of teacher preparation programs under the guidance of Dr. George Noell. An example of one of the reports which lists other alternative certification programs can be found at http://www.regents.doa.louisiana.gov/assets/docs/TeacherPreparation/2010-11ValueAddedAssessmentOverviewofResultsNov212011.pdf.

[8] In most instances this occurs either when teachers have experience from another state, or if teachers were brought in to fill emergency positions in the wake of Hurricane Katrina.

While PTs are expected to secure a job in their intended area of certification, they may change their certification area if they are offered a position in an area different from that of their original intention.

Concurrent with their enrollment in a partner program, PTs must meet several additional criteria to earn their practitioner license. PTs must complete individual requirements for each of the separate partner programs during their year of participation; pay between $3,500 and $4,000 in tuition prior to commencing their practitioner teaching; participate in Teaching for Results, an intensive professional development seminar during their practitioner year; and get an acceptable rating on the summative Performance Assessment System (PAS).[9] Successful completion of these requirements leads to a Level 1 teaching certificate in the state of Louisiana, which is valid for three years.

Participants in the TNTP Academy - Louisiana program are ultimately given a final PAS rating that is the result of a holistic judgment of the many elements that comprise PAS and which are described in greater detail below. The PAS score is based on a portfolio of materials that candidates submit toward the end of their practitioner year of teaching with TNTP Academy - Louisiana. Candidates must pass a minimum threshold on their final PAS score (a score of "promising" or better is passing) in order to continue to be eligible for teacher certification in Louisiana. The PAS is not used as an ongoing form of evaluation and is used only as a way to impart a Level 1 teaching certificate in Louisiana at the conclusion of the year-long TNTP Academy - Louisiana process administered by TNTP.

[9] As of the 2011-2012 school year teachers must also demonstrate effectiveness based on student achievement data. We mention this here for completeness, but this does not apply to the years of data under review in this report.

II. Performance Assessment System (PAS)

A. Components of the PAS Portfolio

PROVIDED THAT A PT meets all other expectations, the receipt of a Level 1 teaching certificate hinges on the candidate receiving a passing score on the PAS portfolio. The PAS portfolio consists of several components that are submitted toward the end of their year as a practitioner teacher. Practitioner teachers submit the portfolio materials to be rated by assessors, who are experienced educators hired and trained by TNTP. Each portfolio is assigned two assessors. The two assessors must agree on a PTs final rating, otherwise a third (or on rare occasions fourth) assessor is assigned to score the portfolio. There are five key pieces of evidence in the portfolio that teachers must submit and on which they are assessed according to a detailed rubric. The five components of the PAS portfolio are:

- a written instructional unit,
- a videotaped lesson,
- an observation of their teaching,
- stakeholder surveys, and
- a student achievement report.

The *instructional unit* must include an outline of the unit, consecutive daily lesson plans, progress reports documenting how students performed on the unit, and student work samples. PTs submit lesson analyses in which they critically reflect on their lesson implementation.

PTs also submit a *videotaped lesson*. Teachers have a choice of which lesson they choose, but it must be one that is documented in their instructional unit plan. In the videotape submission, teachers first give a tour of their classroom to provide context and then teach for thirty-five minutes.

PTs submit a completed *observation* form, based on an on-site observation that is conducted by a program director associated with their partner program.[10]

Program directors are former educators. While each partner program hires its own program directors, all

[10] Each partner program gives a different title to those conducting observations of teachers; TNTP applies the more general moniker program director. Programs directors at teachNOLA are called intervention specialists, at Teach Baton Rouge they are observation specialists, and Teach For America calls them program directors. Program directors are trained separately by each of their respective programs.

fill out one of two common observation forms.[11] The forms have been developed to provide teachers useful feedback on their instructional practice.[12] PTs are observed at least twice in their first year (typically once per semester) for about thirty to fifty-five minutes each time and can choose which of the observation forms to submit as part of their PAS portfolio.[13] To satisfy the observation requirements PTs must also submit a reflective narrative in response to their observation. PT observations do not necessarily occur during one of the lessons included in a PT's instructional unit.

External surveys that are administered to a practitioner teacher's principal, parents, and students are also included in the PAS portfolio. Each teacher distributes and collects surveys from her principal and from a subset of parents and students. The surveys are designed to elicit feedback on teaching performance. The surveys differ slightly depending on the stakeholder and are included in Appendix A.[14] Teachers have some discretion over which students and parents complete the surveys. Teachers in grades K-3 must choose six students to complete the survey, whereas those teaching in grades four through twelve must choose at least one full class of students to complete the student survey. Ten parent surveys are required, but teachers have full discretion over who are asked to complete them. PTs are encouraged to waive their right to review the completed surveys, which are completed in April or May, and all survey responses must be submitted as part of the portfolio.

PTs must also submit a *Demonstrated Student Achievement (DSA) Report*. In this report teachers describe and document evidence of improved academic achievement of their students. Teachers are free to set their own goals as well as to determine and administer diagnostics of their choosing to measure achievement of those goals. These reports do not include state standardized test scores, and are not standardized objective measures of student

---

[11] Program directors have a choice of two forms when performing the observation. The forms differ slightly in that the form favored by TFA is two-pages long and allows for more written feedback. The first page of the two-page form, however, is identical to the other one-page form, and the two-page form places more structure on the feedback than is provided by the one-page version. Both forms are listed in Appendix E. Either form can be used by program directors from any of the three partner programs.

[12] Teachers are encouraged to develop Action reports, which are defined by TNTP as a brief written report to "demonstrate to assessors what specific suggestions your program director offered you and how you used that feedback to influence your classroom practice" (TNTP PAS Handbook, 2009, p. 49).

[13] In some instances partner programs and/or program directors can elect to observe the teacher more than twice. A form is completed for all observations, however, and the teacher can choose among all of them when deciding which to submit.

[14] A slight change was made to student surveys during the years during which we are studying the PAS. These changes were made when TNTP began certifying early childhood teachers. Originally, there were two surveys (one from K-3 and another 4-12), but these were revamped to encompass prek-3 so that they would be appropriate for younger students.

achievement. Rather, achievement reports are based on judgments made by each individual teacher about the progress of his or her students as measured by formative assessments that either the teacher develops or adapts from a teaching resource. Teachers use this information to gauge student understanding and progress on the learning goals that they set at the start of their practitioner teaching year.

Finally, in addition to submitting their portfolio for review, all PTs must receive a passing grade in their content seminar (now called the *Teaching for Results Seminar*). The seminar leader submits an end-of-year report on teacher participation and performance in the year-long content seminar. This end-of-year report is called the Content Seminar Leader Form, or CSL Form. PTs that enter the TNTP Academy - Louisiana program at the start of the year must attend the seminar twice per month, and each meeting is a three-hour professional development seminar in the area where they are seeking certification (e.g., middle school mathematics). In addition to attending the seminars, teachers must also complete some assignments and assessments related to their content area in order to receive a passing grade. Three of the approximately eighteen seminar sessions are devoted to completing the elements of the PAS portfolio.[15]

B. PAS Scoring Overview

THE COMPONENTS of the practitioner teacher's portfolio are scored in five broad "framework areas" that TNTP uses as the overall structure of the PAS. The framework areas are largely analogous to the Framework for Teaching developed by Charlotte Danielson, which is a classroom observation system used by many school districts in the U.S. The five framework areas are:[16]

- instructional design and delivery,
- classroom environment,
- assessment,
- professionalism, and
- student achievement.

[15] For more information about the content seminar see http://tntpacademylouisiana.ttrack.org/CurrentParticipants/TeachingforResultsSeminars.aspx

[16] Information on the Danielson Framework can be found in Danielson (1996).

Within each framework area there are subscores called "critical elements," and each of these critical element scores are generated by assessors based on the components that the PTs submit as part of their portfolio. As an example, Figure 1 depicts the rubric used to score the framework area Professionalism (this is an excerpt from the overall Global Rubric in Appendix B). Each of the rows in Figure 1 depicts the three critical elements that are included in Professionalism, and each of the columns represents a portfolio component that is used to judge a particular critical element. Within this matrix of rows (critical elements) and columns (components) some "cells" are shaded as gray if that particular critical element does not require that portfolio component to create a critical element score. For instance, in the first column Figure 1 illustrates that the videotape component is used to generate critical element scores for the first two critical elements in Professionalism, but not the third.

Critical element scores are determined by first scoring each of the cells in the row that corresponds to a particular critical element for a specific framework area. For example, the first critical element in Professionalism, "Reflects on and revises practice continuously to improve teaching performance," consists of scores on both videotape and PD observation components. As is shown in the first cell of this row, the videotape component must provide evidence that "Lesson Analysis reflects critically on teaching strengths and areas for improvement as shown in the videotaped segment." After reading this statement, assessors are expected to rate the PT on this component within this critical element. Assessors must choose a score from; "Ex" for exemplary, "E for effective, "P" for promising, and "I" for ineffective.

FIGURE 1. PROFESSIONALISM PORTION OF THE PAS GLOBAL RUBRIC

PAS Rubrics: Professionalism

CRITICAL ELEMENT	VIDEOTAPE	SURVEYS	PD OBSERVATIONS	CSL FORM
Reflects on and Revises practice Continuously to Improve teaching performance	☐ Lesson Analysis reflects critically on teaching strengths and areas for improvement as shown in the videotaped segment. Ex E P I		☐ Action Report reflects critically on PD feedback and describes specific actions or steps taken to improve classroom performance. Ex E P I	
Establishes positive, professional relationships with student, parents, and colleagues	☐ Teacher maintains positive rapport with students without compromising role as instructional leader. Ex E P I	☐ PARENTS: "Always" / "Usually" as a majority of responses to question 6 ☐ PARENTS: Most parents answer "yes" to ☐ PRINCIPAL: "Agree/ Somewhat Agree"as a response to question 8 Ex E P I		
Fulfills professional responsibilities				☐ Teacher successfully meets seminar requirements and expectations (the "pass" box is checked). Ex E P I

For each of the five Framework areas, scores given by assessors in each of these "cells" on the Global Rubric are then transferred to the Global Synthesis Form where they are aggregated up to overall Framework scores. Figure 2 depicts the portion of the Global Synthesis form relating to the Professionalism framework (for the complete Global Synthesis Form see Appendix C). The Global Synthesis Form has five rows, one for each of the non-gray cells depicted in the Professionalism portion of the Global Rubric. Scores from each of the cells on the Global Rubric are transferred to the "Rating" column in the corresponding row of the appropriate framework area. For instance, the score from the videotape component for the first critical element under Professionalism would be transferred to the first row of the "Rating" column in Figure 2. Next to each rating entry for each component

score, the Global Rubric specifies an approximate "weight" that should be afforded to the component score for that critical element. The approximate weight is given in the column headed as "Sig." for significance, with possible weights L, M, and H, which correspond to low, medium, and high, respectively. In the first row for Professionalism, the score on the videotape component for the first critical element is meant to get "medium" weighting within the Professionalism framework. To create an aggregate framework score from each of these component scores within critical elements, the assessors are meant to apply a holistic approach, rather than simply averaging each of the scores within this framework area (note that the weight, H is not used in Professionalism). Using each of the five scores entered in the Global Synthesis Form for Professionalism, the assessor generates one overall score for professionalism on the same scale described above (Ex, E, P, and I).

FIGURE 2. PROFESSIONALISM PORTION OF THE GLOBAL SYNTHESIS FORM

PAS Synthesis Form

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Reflects on and revises practice continuously to improve teaching performance	Videotape		M	Ex	PROFESSIONALISM
	PD Observations		L	E	
Establishes positive, professional relationships with students, parents/ guardians, and colleagues	Videotape		M	P	
	Surveys		M		
Fulfills professional responsibilities	CSL Form		L	I	

Once the Global Synthesis Form is completed for each of the framework areas the scores are transferred again to the PAS Final Rating Form, which is depicted in Figure 3.

FIGURE 3. PAS FINAL RATING FORM

1. Framework Area Ratings

Transfer the Framework Area ratings from Synthesis Forms pag 1-4 below:

FRAMEWORK AREA	WEIGHT	RATINGS (Ex, E, P, I)
Classroom Environment/Culture	25%	
Instructional Design & Delivery	25%	
Assessment	15%	
Professionalism	15%	
Student Achievement	20%	

2. Final Rating

Using your professional judgement and the weights above, assign a final rating to this teacher's overall performance in this portfolio.

EXEMPLARY EFFECTIVE PROMISING INEFFECTIVE

Scores from each of the framework areas on the Global Synthesis Form are to be translated to the corresponding rows on the Final Rating Form. Despite their numeric values, the weights shown in the "Weight" column in Figure 3 are only approximate weights, and assessors are not instructed to simply take a weighted average of the numeric values of each of their framework scores.[17] In fact, as shown in the bottom portion of Figure 3, assessors are supposed to choose a Final Rating using "their professional judgment and the weights…." Not shown in Figure 3 is a final portion of the Final Rating Form, where assessors are expected to write a paragraph explaining the holistic judgment they applied to generate the final rating.

The final portfolio rating that a practitioner teacher receives—the one that determines whether they pass and are recommended to receive a license issued by the state or not—is based on a consensus score from the final ratings from each of the two originally assigned assessors. If the two assessors agree on their final rating, then the teacher receives the agreed upon grade. If the assessors differ in their final rating, then the portfolio is rated by a third assessor, and, if necessary, a fourth. A passing grade is one that is "promising" or higher (a score of two through four on the four-point scale), while teachers who receive a final rating of "Ineffective" (score of one) do

[17] Note that the analyses below take several approaches to incorporating each of the scores that an assessor is asked to record. In some cases the scores are used just as they were recorded by the assessors, and in others the scores recorded are averaged in the ways described in the Analysis section of this report.

not pass. Only the final portfolio rating is used to make the decision about passing or failing. Teachers that receive a passing score receive Level 1 licensure, and those that fail may be eligible to participate in the TNTP Academy - Louisiana process in the subsequent school year.

Though the process of assessing components of the PAS portfolio (described above) is holistic in nature, there is a well-defined rubric designed to help assessors make appropriate assessments of the quality of the portfolio components. Assessors are trained on this rubric and have them available during the scoring process. The rubric used in training assessors contains some examples of how to distinguish between rating scores (e.g., how to distinguish Exemplary from Effective, and/or Promising from Ineffective). For instance, to continue with the example from the Professionalism framework area described above, Figure 4 depicts the assistance given to assessors when deciding which score to apply to the videotape component of the two critical elements within Professionalism.

FIGURE 4. VIDEOTAPE COMPONENT OF PROFESSIONALISM FRAMEWORK IN THE ANALYTIC RUBRIC

CRITICAL ELEMENT	KEY INDICATOR	EVIDENCE/NOTES	RATING
Reflects on and revises practice continuously to improve teaching performance	☐ Lesson Analysis reflects **critically** on teaching strengths and areas for improvement as shown in videotaped segment. *P: Lesson Analysis critically describes strengths, but not areas of improvement in teaching performance* *Ex. Lesson Analysis identifies areas for improvement, and lists specific resources and a plan the teacher will utilize to develop these areas.*		Ex. B P I
Establishes positive, professional relationships with students, parents/ guardians, and school colleagues	☐ Teacher maintains positive rapport with student **without compromising** role as instructional leader.		Ex. B P I

In the top row of Figure 4, in the "Key Indicator" column, assessors receive guidance on how to differentiate "promising" scores, and "exemplary" scores for the first critical element within Professionalism. Presumably, if assessors find that the component does not rise to the level of exemplary, but exceeds promising, then the PT receives a score of effective. If, however, the component is not worthy of the promising score, then the PT receives a score of ineffective on this component of the first critical element of Professionalism. As is evident in the second row of Figure 4, not all guidance in this rubric is well-defined. The guidance given within other individual portions of the Analytic Rubric varies in ways similar to what is depicted in Figure 4.

III. Data and Sample

DATA FOR THE STUDY was provided by both the Louisiana Department of Education (LDOE) and TNTP. TNTP provided PAS evaluation data along with some teacher background information, while the LDOE contributed student class assignment, test score, and demographic data. These files were merged to create a student-teacher level file, the details of which are further described below.

A. Louisiana Department of Education Data

The Louisiana student enrollment information files were structured at the teacher-class-student level, and contain student enrollment information for teachers in each of the four TNTP Academy - Louisiana cohorts included in this evaluation, school year 2005-06 through school year 2008-09. The data also include student test scores from the 2003-04 through 2009-10 school years, for the LEAP, iLEAP, GEE, LAA1, LAA2, and the Iowa Test of Basic Skills (ITBS) assessments.[18] For the purpose of evaluating the PAS, only the LEAP and iLEAP scores for regular education students are used because these assessments have been consistently implemented in the policy window under study here. In addition to the enrollment and test score data, the Louisiana data files provide a set of student demographics which includes variables for a student's age, sex, ethnicity, free or reduced-price lunch status, English Language Learner status, dates of enrollment, and absences.

[18] Testing takes place in grade 3 through 11 in Louisiana, but the subjects tested varies by grade in school. In grades 4 and 8 the test given is the Louisiana Educational Assessment Program (LEAP). This test is used for NCLB accountability purposes and is also used to make decisions about whether to retain students in grade. In grades 3, 5, 6, 7, and 9 the Integrated Louisiana Educational Assessment Program (iLEAP) is the test of record, and in grades 10 and 11 the Graduation Exit Examination (GEE) is used. All tests are placed on the same scale of 100-500. It is also important to note that approximately 1% of all students can be tested using either form one or two of the Louisiana Alternative Assessment (LAA1 or LAA2), which is an alternative assessment for students with more severe disabilities.

B. TNTP Data

Data that we use from TNTP includes the teacher-level scores on their PAS portfolios for the four cohorts enrolled in the TNTP Academy - Louisiana during the years 2004-05 through 2008-09.[19] This data includes each teacher's final portfolio score on PAS, scores from the five framework areas (one for each of the portfolio assessors) and scores on the individual pieces of evidence within the Critical Elements that assessors evaluated. The TNTP data also includes some teacher background variables including: teacher's partner program, degree information, degree-granting institution, GPA, standardized test scores required for teaching certification, program placement, and anticipated area of teaching certification. Missing teacher background data limit their inclusion in these analyses.

C. Linking and Limitations of the Data

Names of TNTP teachers with PAS scores in each cohort were sent to the LDOE to match with its database. Unfortunately, all teachers in the cohorts were not matched to LDOE files. We present in Table 1 the number of teachers in our analytic sample who had PAS scores and the number of those teachers that the LDOE was able to match, by cohort.

TABLE 1. NUMBER OF TEACHERS BY COHORT MATCHED TO LOUISIANA DEPARTMENT OF EDUCATION DATA.

COHORT	WITH PAS SCORES	MATCHED BY LA DOE	PERCENT MATCHED
2005-06	78	75	96%
2006-07	90	84	93%
2007-08	244	207	84%
2008-09	358	292	81%
TOTAL TEACHERS	770	658	85%

[19] While we have data available for the 2009-10 cohort, the final portfolio scoring for this cohort is different than for the earlier cohorts. Teachers in the four earlier cohorts received integer scores of 1, 2, 3, or 4 for their final portfolio score whereas teachers in the 2009-2010 cohort were assigned a dichotomous pass or fail portfolio score. Form comparability across years we therefore only use the four earlier cohorts in our analysis. See Appendix A for further explanation.

All cohorts of TNTP teachers were less than perfectly matched with LDOE data. This imperfect match is evidenced by comparing columns (1) and (2) in Table 1, and is summarized by the matching rate in column (3). The imperfect match rate is a potential threat to the external validity of our findings. If teachers who were matched in the LDOE data are systematically different than those who do not match, then the results of this analysis will only generalize to the universe of teachers who are similar to those in the matched sample, rather than the overall universe of teachers.

Data were additionally limited by the fact that not all teachers taught subjects that were tested in Louisiana, and that the LEAP and iLEAP exams are only administered in grades three through eleven. Because we use prior year test scores as a control variable in our analyses, we limit our analytic sample to grades four through eleven. Furthermore, the LEAP and iLEAP exams are administered differently by subject. The result is that our analyses include grades four through nine in mathematics, English Language Arts (ELA), and reading, while social studies and science include only grades four through eight. Any teachers who taught in untested elementary grades or specialty subjects (e.g., music, foreign language, art) are excluded from our analytic sample. To simplify analyses, we examine teachers within the same subject area, but across grades, as separate samples.

It is worth noting that all studies seeking to validate a comprehensive system of teacher evaluation struggle with the reality that these measures can be linked to student test scores in a limited number of subjects. This data from TNTP and the State of Louisiana is notably different in that it includes five tested areas—mathematics, ELA, reading, science, and social studies—which is more than what exists in many systems.

D. Analytic Sample

The analytic sample includes teachers with PAS scores (scored on the same scale), LDOE student course data, and test scores for sufficient number of students to make a value-added calculation. Within subject area analyses we limit our sample to those classes whose enrollment is less than 50% students with disabilities, and for whom more than half the students have prior test scores in the subject area of interest.[20] We also only include classes that contain five or more students.

[20] Missing scores for the other students are imputed as described below in the empirical strategy section.

Table 2a displays several characteristics of the overall sample of teachers and students included in the analyses that follow. Each column in Table 2a corresponds to a different tested subject and therefore includes a different subset of teachers. Panel A of the Table 2a summary statistics—including baseline test scores, demographic information, and grade of enrollment—are presented for the students included in each of the analyses. Student demographics are relatively similar across the five samples. The characteristics of teachers included in each sample are shown in Panel B, where teacher gender and the percentage who have attained an advanced degree are listed for each subject area, as are the counts of teachers in a given subject area from each of the four TNTP cohorts. The practitioner teachers in this sample are largely female, and as might be expected with early career teachers, few have an advanced degree.

Mean PAS scores received by teachers in each sample are displayed in Panel C of Table 2a. In this panel, rows one through five show end-of-year framework scores, while row six gives the mean PAS portfolio score for teachers in each sample. Framework area scores average in the 2.5 to 3 range (on the 1-4 scale) with professionalism scores consistently averaging higher than the other four framework areas. Within each framework, mean scores are similar across the sample, as is the case with final portfolio scores.

TABLE 2A. OBSERVABLE STUDENT AND TEACHER CHARACTERISTICS OF ESTIMATION SAMPLE

	Math	ELA	Reading	Science	Social Studies
Student Characteristics					
Baseline LEAP Score	-0.576 (1.01)	-0.58 (1.032)	-0.648 (1.071)	-0.499 (0.998)	-0.509 (1.052)
Male	50.5%	50.7%	51.8%	50.9%	52.4%
Racial/Ethnic Minority	89.9%	87.6%	85.3%	90.0%	91.5%
White	10.1%	12.4%	14.7%	10.0%	8.5%
Special Education	9.3%	9.5%	11.9%	10.8%	10.7%
English Language Learner	3.2%	3.1%	2.3%	3.3%	2.2%
Free or Reduced Lunch	87.2%	87.7%	88.4%	91.4%	90.0%
Retained in Grade	12.1%	10.4%	13.4%	9.2%	9.5%
Grade 4	5.3%	8.4%	12.3%	10.2%	20.1%
Grade 5	8.1%	13.1%	10.3%	10.3%	20.1%
Grade 6	24.4%	21.3%	21.0%	25.1%	26.9%
Grade 7	24.8%	25.0%	17.5%	35.7%	15.7%
Grade 8	21.1%	17.5%	25.3%	18.7%	17.2%
Grade 9	16.3%	14.7%	13.5%	0.0%	0.0%
Total Student Observations	4732	4030	2271	4067	2146
Teacher Characteristics					
Male	41.4%	27.0%	31.2%	28.4%	42.2%
Advanced Degree	3.9%	6.0%	2.0%	2.6%	2.6%
2005 Cohort	2	1	0	3	2
	2.4%	1.7%	0.0%	5.0%	9.5%
2006 Cohort	17	11	0	6	2
	23.4%	11.6%	0.0%	8.9%	1.4%
2007 Cohort	43	38	20	31	17
	35.4%	38.4%	32.1%	32.3%	20.6%
2008 Cohort	47	56	44	43	38
	38.8%	48.3%	67.9%	53.8%	68.6%
Total Teacher Observations	109	106	64	83	59
Framework Area Scores					
Classroom Environment	2.823 (0.49)	2.735 (0.513)	2.763 (0.53)	2.705 (0.488)	2.809 (0.54)
Instructional Design & Delivery	2.836 (0.357)	2.739 (0.424)	2.78 (0.449)	2.877 (0.375)	2.857 (0.414)
Assessment	2.657 (0.405)	2.532 (0.48)	2.509 (0.553)	2.673 (0.458)	2.645 (0.384)
Professionalism	2.99 (0.331)	2.95 (0.437)	2.973 (0.454)	3.018 (0.393)	3.061 (0.426)
Student Achievement	2.769 (0.425)	2.696 (0.486)	2.726 (0.506)	2.73 (0.412)	2.816 (0.449)
Final Portfolio Rating	2.882 (0.357)	2.749 (0.466)	2.778 (0.481)	2.838 (0.399)	2.774 (0.476)

Note: Descriptive statistics include grades 4-9 students of TNTP teachers with PAS portfolio ratings from 2005-06 through 2008-09 in classes with at least 5 students and less than 50 percent special education students. Of students retained in grade, most repeat 4th and 8th grades due to Louisiana's test-based retention policy.

Table 2b gives the pair-wise correlations for framework area scores as given by evaluators. These correlations range from 0.45 for the correlation between professionalism and student achievement to 0.58 between instructional design and delivery and student achievement. We note that these correlations are relatively low compared to correlations that Kane et al. (2011) found in Cincinnati across the eight standards found in the classroom environment and classroom instruction domains of the Danielson Framework. In that study the pair-wise correlations among the eight standards ranged from 0.619 to 0.813, with most correlations closer to the higher end.

TABLE 2B, PANEL A. CORRELATION MATRICES OF FRAMEWORK AND CRITICAL ELEMENT SCORES AS GIVEN BY EVALUATORS.

	Final Classroom Environment	Final Instructional Design & Delivery	Final Student Assessment	Final Professionalism	Final Student Achievement
Classroom Environment	1				
Instructional Design & Delivery	0.589	1			
Assessment	0.460	0.709	1		
Professionalism	0.638	0.605	0.504	1	
Student Achievement	0.573	0.668	0.658	0.595	1

Note: Additional correlation tables are included at the end of this report.

Table 2c presents information on the inter-rater reliability (IRR) of the PAS evaluators. While there is much information to peruse in this table, two general observations emerge. The first is that the IRR levels are relatively low for all of the measures. There is only one instance, Positive Environment: Survey within the Classroom Environment framework, where the evaluators were in agreement at least 70% of the time. Furthermore, the highest Kappa measures are around 0.45 (in three different cases) and the highest Spearman's rho is around 0.57 (in two instances). These figures are below what are generally considered acceptable levels of IRR.[21]

The second observation is that the highest levels of agreement tend to be in the use of the various surveys. While not always the case, it is usually the case that within the framework, one sees the highest IRRs associated with surveys.

TABLE 2C. PERCENT AGREEMENT AMONG EVALUATORS BY CRITICAL ELEMENT AND SOURCE

	Percent Agreement Among Evaluators	Inter-Rater Reliability – Weighted Cohen's Kappa	Inter-Rater Reliability – Spearman's Rho
Final Portfolio Rating	98.8%	0.35	0.439
Final Class Environment Rating	60.5%	0.321	0.363
Final Instructional Design & Delivery Rating	58.7%	0.236	0.303
Final Assessment Rating	51.2%	0.254	0.316
Final Professionalism Rating	65.0%	0.344	0.404
Final Student Achievement Rating	58.2%	0.495	0.549
Classroom Environment			
Systems & Routines			
Video	52.7%	0.306	0.389
Observation	57.6%	0.291	0.363
Communicates Standards			
Video	47.8%	0.332	0.412
Survey	66.4%	0.43	0.517
Observation	57.3%	0.365	0.45

[21] Landis, J.R. & Koch, G.G. (1977). "The measurement of observer agreement for categorical data." Biometrics 33(1): 159–174.

TABLE 2C. CONTINUED

Physical Environment			
Video	61.6%	0.305	0.351
Positive Environment			
Video	50.0%	0.303	0.356
Survey	70.1%	0.297	0.351
Observation	56.0%	0.283	0.35
Instructional Design & Delivery			
Backwards Plan			
Instructional Unit	58.1%	0.115	0.122
Survey	66.5%	0.291	0.343
Observation	51.6%	0.241	0.302
General Student Instruction			
Instructional Unit	58.2%	0.139	0.157
Video	52.2%	0.349	0.411
Survey	54.9%	0.419	0.531
Observation	51.2%	0.197	0.245
Modifies Instruction			
Instructional Unit	51.6%	0.258	0.339
Video	55.9%	0.304	0.382
Survey	58.6%	0.428	0.542
Communicates Progress			
Instructional Unit	59.2%	0.344	0.377
Survey	48.7%	0.282	0.357
Observation	24.0%	0.167	0.257
Professionalism			
Reflects on Practice			
Video	54.4%	0.21	0.237
Observation	59.2%	0.211	0.253
Positive Relationships			
Video	56.6%	0.301	0.342
Survey	67.3%	0.222	0.237
Fulfills Responsibilities			
CSL Form	64.6%	0.319	0.346

TABLE 2C. CONTINUED

Student Achievement			
Evidence Progress			
Instructional Unit	52.2%	0.253	0.322
Video	52.7%	0.303	0.356
DSA	50.7%	0.278	0.359
Survey	65.9%	0.319	0.406

Note: The percent agreement among evaluators represents the percent of teachers in the sample for whom all evaluators assigned the same ranking for a given critical element.

In the analyses that follow, we will be using PAS scores based on the evaluators final ratings, and we will also be using PAS scores based on information that exists at the critical element level. Table 2d gives summary statistics on both of these measures for each of the different subject-based samples and by framework area. Mean framework scores from the evaluators are given in Panel A of Table 2d, and constructed mean framework scores are given in Panel B of the table. To obtain the measures in Panel B we first averaged all critical element scores across evaluators and across evidence source within-teacher and within-framework. The mean scores in Panel B are then the average across teachers of these constructed framework scores. Looking across the information in the two panels, mean constructed scores are quite similar to the means scores based on evaluator end-of-year ratings. Despite the similar summary measures, however, it is an open question as to whether evaluators' end-of-year ratings or the constructed measures are better predictors of teacher effectiveness. This is one of the questions we pursue below.

TABLE 2D. SUMMARY STATISTICS OF AVERAGE FRAMEWORK AREA SCORES AS GIVEN BY EVALUATORS AND CONSTRUCTED FROM THE AVERAGE OF ALL CRITICAL ELEMENTS ACROSS ALL SOURCES AND ASSESSORS, BY SUBJECT AREA ANALYSIS SAMPLE.

Panel A. Mean and standard deviation of framework area scores as given by evaluators by subject area analysis sample

Framework Area	Math	ELA	Reading	Science	Social Studies
Classroom Environment	2.823	2.735	2.763	2.705	2.809
	(0.49)	(0.513)	(0.53)	(0.488)	(0.54)
Instructional Design & Delivery	2.836	2.739	2.78	2.877	2.857
	(0.357)	(0.424)	(0.449)	(0.375)	(0.414)
Assessment	2.657	2.532	2.509	2.673	2.645
	(0.405)	(0.48)	(0.553)	(0.458)	(0.384)
Professionalism	2.99	2.95	2.973	3.018	3.061
	(0.331)	(0.437)	(0.454)	(0.393)	(0.426)
Student Achievement	2.769	2.696	2.726	2.73	2.816
	(0.425)	(0.486)	(0.506)	(0.412)	(0.449)

Panel B. Mean and standard deviation of constructed framework area scores by subject area analysis sample

Framework Area	Math	ELA	Reading	Science	Social Studies
Classroom Environment	2.829	2.744	2.778	2.755	2.843
	(0.358)	(0.404)	(0.43)	(0.339)	(0.398)
Instructional Design & Delivery	2.814	2.749	2.781	2.817	2.852
	(0.253)	(0.319)	(0.37)	(0.272)	(0.3)
Assessment	2.685	2.617	2.603	2.679	2.717
	(0.263)	(0.329)	(0.403)	(0.305)	(0.271)
Professionalism	2.99	3.005	3.03	3.024	3.076
	(0.305)	(0.349)	(0.366)	(0.294)	(0.347)
Student Achievement	2.721	2.702	2.686	2.703	2.771
	(0.364)	(0.373)	(0.455)	(0.364)	(0.385)
Total Teacher Observations	109	106	64	83	59

Standard errors in parentheses.
Note: Sample includes grades 4-9 teachers with PAS portfolio ratings from 2005-06 through 2008-09 in classes with at least 5 students and less than 50 percent special education students.

IV. Empirical Strategy

WE ARE INTERESTED in estimating the relationship between teaching practices as measured by the PAS evaluation tool and teacher effectiveness in promoting student achievement growth as measured by student test scores on the aforementioned Leap and iLeap tests. We model the relationship between student achievement and PAS scores with equation 1:

$$(1)\quad a_{i,j,k,t} = PAS_j\gamma + A_{i,t-n}\alpha + S_{i,t}\beta + P_{j,k,t}\delta + \rho_t + \varepsilon_{i,j,k,t}$$

where $a_{i,j,k,t}$ is the achievement score of student i taught by teacher j in school k in year t. Each teacher j appears only once in our data, in year *t* when they were first evaluated with the PAS tool. To isolate the relationship between PAS scores and student achievement growth, equation 1 controls for prior student achievement $A_{i,t-n}$, other observable characteristics of the student $S_{i,t}$ and her peers $P_{j,k,t}$. Equation 1 also includes grade by year fixed effects, ρ_t.

The vectors $A_{i,t-1}$, $S_{i,t}$, and $P_{j,k,t}$ include the following variables:

- $A_{i,t-n}$, a vector of information regarding a student *i*'s prior achievement, includes:
    - o $a_{i,t-n}$ student *i*'s test score in the same subject (e.g., math when predicting math) the previous school year *t-1* and the year before that *t-2*,[22]
    - o the square and cube of $a_{i,t-1}$ and $a_{i,t-2}$,
    - o the interactions of $a_{i,t-1}$ and $a_{i,t-2}$ with the grade level of that test,
    - o the interactions of $a_{i,t-1}$ and $a_{i,t-2}$, the grade level of that test, and an indicator for being a grade repeater, and
    - o $a'_{i,t-1}$ student *i*'s test score in the different subjects (e.g., reading, science, and social studies when predicting math) the previous school year *t-1* and the year before that *t-2*.

[22] In most cases, this will be the student's test score in the previous grade. For students who are repeating the grade, however, this will be the student's score in the same grade. As noted in the text, we standardize test scores at the grade x subject x year level.

Sometimes a small number of students in a class will not have taken one or more exams the previous year, and thus do not have values for $a_{i,t-1}$ or $a'_{i,t-1}$ or both. If a student is missing data for prior tests in any subject, we impute the score as zero, and include an indicator variable in $A_{i,t-1}$ identifying such cases.

All test scores, $a_{i,t-n}$ and $a'_{i,t-n}$ along with $a_{i,j,k,t}$, are standardized with a mean of zero and a standard deviation of one at the subject x grade x year level of the test.[23] This standardization is calculated based on the student's scaled score compared to other students in the state who took the same test in the same grade and year.

- $S_{i,t}$, a vector of other observable characteristics of student *i* during school year *t*, includes indicator variables for:
    - gender,
    - each racial or ethnic subgroup,
    - each free or reduced-price lunch program classification (as a proxy for family income),
    - English Language Learner classification,
    - an indicator for special education,
    - whether the student was new to the school (this includes structural transitions from elementary to middle schools, and non-structural changes), and
    - an indicator for substantial absence in the prior year (i.e., greater than 10% of enrolled days in year *t-1*).
- $P_{j,k,t}$, a vector of observable characteristics of student *i*'s peers in class *j* and peers in the same grade level at the student's school, includes separately for class *j* and the school grade-level cohort:
    - the means of the elements of $S_{i.t}$,
    - the means and standard deviations of $a_{i,t-n}$ and each of the tests in the $a'_{i,t-1}$ vector,
    - the proportion of peers who are missing test scores for $a_{i,t-n}$ and each of the tests in the $a'_{i,t-1}$ vector, and
    - the number of students in class *j* and the number of students in the grade-level cohort.

The parameter of interest in equation 1 is γ, a measure of the relationship between a teacher's PAS score and student achievement gain. In the analyses that follow, we explore various formulations of PAS from the

[23] We note that the mean baseline test scores of the students in our sample are lower than other students across the state indicating that TNTP teachers are serving a lower performing student population than the average teacher in the state.

most aggregated—a teacher's end-of-year final portfolio score—to other specifications of PAS that incorporate less aggregated information.

Unlike what has been the case in other settings, we find that there is enough independent variation among the five framework scores—with correlations that range from 0.42 to 0.60—that in the relevant instances, the framework scores can serve as appropriate elements of a PAS vector.[24] The goal in exploring different PAS formulations is to not only determine the extent to which the PAS tool may be an effective measure of a teacher's ability to promote student achievement growth, but to also determine whether there are particular elements of PAS that are more predictive of teacher effectiveness than others. Across all different formulations of PAS, we fit equation 1 with ordinary least squares (OLS), clustering standard errors at the teacher level.[25]

[24] Kane et al. (2011) found that the equivalent of the PAS framework scores in the Cincinnati evaluation system, the eight "standard" scores of the Cincinnati evaluation tool, were so highly correlated that it was more meaningful to instead use in the analysis the first three principal components that underlay the standard scores in the Cincinnati data. A similar principal components analysis of the five PAS framework scores identifies one principal component as explaining most of the underlying variation. This first principal component is effectively the equally weighted average of the five framework scores, a measure that we explore in our analysis in tables 4a-4e.

[25] Note that in our analysis, we run models for each subject separately so each student is only associated with one classroom and one teacher (although some teachers have multiple classrooms), so that clustering by teacher subsumes clustering by classroom.

V. Results

A. Overall PAS Ratings as a Measure of Teacher Effectiveness

THE OVERARCHING QUESTION we are exploring is the extent to which the PAS tool distinguishes teachers by their ability to increase student academic achievement. A first approach at answering this question is to determine whether teachers with higher end-of-year PAS portfolio ratings are more effective at raising student test scores. The results in Table 3 are based on estimates of equation 1 where the PAS measure is teacher j's final PAS portfolio score in year t as determined holistically by the PAS evaluators. In Panel A of Table 3, the final portfolio score is constrained to have a linear relationship with achievement, while in Panel B we allow each PAS integer score 1-4 to have a separate relationship to student achievement. Throughout the paper, we fit each model separately by subject. To be clear, the PAS measures in Table 3 are based on the end-of-year single PAS portfolio score for each teacher that reflects the consensus view of all evaluators across all framework scores.

TABLE 3. ESTIMATED RELATIONSHIP BETWEEN STUDENT ACHIEVEMENT GROWTH AND FINAL PAS PORTFOLIO RATINGS FOR 2005-06 THROUGH 2008-09

Panel A. Estimated relationship between final PAS portfolio rating and student achievement growth

	Math	ELA	Reading	Science	Social Studies
Final Portfolio Rating	0.166*	0.0413	0.143*	0.162***	0.247**
	(0.0765)	(0.0403)	(0.0569)	(0.0444)	(0.0864)

Panel B. Difference in relationship between final PAS portfolio ratings and student achievement growth in comparison to final portfolio rating of 2

	Math	ELA	Reading	Science	Social Studies
Final Portfolio Rating of 1	-0.104	0.174	-0.147	0.209	-0.479*
	(0.126)	(0.106)	(0.283)	(0.133)	(0.223)
Final Portfolio Rating of 3	0.170	0.0581	0.132*	0.184***	0.130
	(0.0862)	(0.0457)	(0.0522)	(0.0467)	(0.116)
Final Portfolio Rating of 4	0.339	0.314**	0.526***	0.430	0.920***
	(0.177)	(0.109)	(0.122)	(0.303)	(0.213)
Teacher Observations	109	106	64	83	59
Student Observations	4732	4030	2271	4067	2146
Adjusted R^2	0.489	0.532	0.413	0.422	0.361
Mean of LEAP Assessment	-0.576	-0.580	-0.648	-0.499	-0.509
	(1.010)	(1.032)	(1.071)	(0.998)	(1.052)

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The estimates in Panel A of Table 3 indicate a positive and statistically significant relationship between the final portfolio score and student achievement gain in four of the five subjects. To the extent that results based on equation 1 can be interpreted as causal, estimates in the first column of Panel A indicate that teachers who score one unit higher on the PAS portfolio score would be expected to generate math achievement gains that are 0.17 of a standard deviation higher than teachers who score one unit lower on their portfolio score. To put this in context, an effect size of 0.17 would move a student from the 50th to the 57th percentile in the student math achievement distribution. For reading, a one unit higher portfolio score is associated with achievement gains that are 0.14 of a standard deviation larger, and in science and social studies the estimated achievement gains are 0.16 and 0.25, respectively. The one-quarter standard deviation estimated achievement gain in social studies is associated with moving a student from the 50th to the 60th percentile of the student social studies distribution.

In Panel B of Table 3, we relax the linear constraint on PAS by entering indicator variables for PAS scores of 1, 3, and 4, with a score of 2 being the excluded category. We focus on the comparison between teachers who score 3 versus those whose final portfolio score is 2 because in each of the five samples the great bulk of teachers are in these two categories. For example, Appendix I shows that about 96% of the teachers in the math sample receive scores of 2 or 3, with only one of the 127 teachers having a portfolio score of 1 and only three having a score of 4. This massing of teachers in categories 2 and 3 is also present across the other subjects, so that regardless of the subject, estimates related to PAS portfolio scores will largely be driven by teachers who score either a 2 or 3.

This point is brought home by comparing the point estimates on the category 3 indicator in Panel B to the estimated linear effects in Panel A. Except for social studies, the estimates are very similar within subject across the two specifications in Panels A and B, though only reading and science are statistically significant in Panel B. Thus, based on the variation between teachers who score 3 versus 2, one would conclude that in reading and science higher-scoring teachers generate achievement levels that are around 0.13 to 0.18 of a standard deviation higher in those subjects. We note that the estimates on the category 4 indicator in ELA and social studies, in particular, suggest that while PAS evaluators may be less able to distinguish teacher effectiveness for the bulk of teachers in these subjects than in reading and science, there is some evidence that can identify teachers who are top performers in ELA and social studies. Likewise, the statistically significant estimate on the category 1 indicator in social studies suggests evaluators can also pick out the lowest performing teachers in this subject.

The overarching lesson from Table 3 is that PAS portfolio scores do tend to be predictive of a teacher's

ability to promote student achievement growth in all subjects except for ELA. A logical question to ask is: could one take steps that would make PAS-based evaluation even more predictive? One could consider options such as increasing and improving the training of evaluators, or perhaps altering the scoring rubric. We ask a more fundamental question: are there ways to utilize the information collected by the current version of PAS that result in greater predictive capacity?

One issue with the current practice of using an end-of-year holistic assessment to group teachers into one of four categories is that this method ignores a substantial amount of within-category variation in teacher effectiveness. This variation is evident in Figures 1a-1e (shown in Additional Tables and Figures) that graph the value-added scores of teachers in each of the subject area samples against their final portfolio rating.[26] As a first observation, the fitted regression lines in these graphs display approximately and graphically the estimates from Panel A of Table 3. The additional and important information that comes from the figures, however, is that while, on average, teachers assigned a 3 on their final portfolio tend to outperform teachers assigned a 2, there is also a considerable amount of variation in effectiveness within each of the two categories in every subject area sample.

One way to capture some of this variation is to use all of the available information contained in the critical element-level scores. A simple way to do this is to create an overall PAS score based on the within-teacher average of all critical element scores across all evaluators. While this might not be the optimum weighting of critical element scores across frameworks and sources, it is a logical first step.

Consider the results shown in Table 4a. This table presents estimates from five different models fit on the math sample. In the first column, model 1, we reproduce the math estimate from Panel A of Table 3, the linear specification of the PAS holistic-determined portfolio scores. The estimate in the second column for model 2 is from a specification where a teacher's PAS measure is the within-teacher average of all of the critical element scores. Comparing estimates between the two models indicates that a one-unit increase in the constructed PAS score relates to roughly twice the student achievement gain as a one-unit increase in the holistic-assigned portfolio score, i.e. moving from a 2 to 3 in that score.

[26] Value-added scores were constructed as follows: 1) omitting the PAS variable, equation 1 was fit and the residuals were captured; 2) the within-teacher residuals were then averaged to construct each teacher's value-added score.

TABLE 4A. COMPARISONS OF THE MATH ACHIEVEMENT-PAS RELATIONSHIP ACROSS MODELS

	Model 1	Model 2	Model 3	Model 4	Model 5
Final Portfolio Rating	0.166* (0.0763)				
Average of All Critical Element Scores		0.336*** (0.0886)			
Final Portfolio Rating of 1			-0.104 (0.126)		
Final Portfolio Rating of 3			0.170 (0.0862)		
Final Portfolio Rating of 4			0.339 (0.177)		
Average Critical Element Score Rank of 1				-0.162 (0.106)	
Average Critical Element Score Rank of 3				0.304*** (0.0769)	
Average Critical Element Score Rank of 4				0.682*** (0.133)	
Teacher Effect Rank of 1					-0.169* (0.0750)
Teacher Effect Rank of 3					0.426*** (0.0442)
Teacher Effect Rank of 4					1.200*** (0.117)
Teacher Observations	109	109	109	109	109
Student Observations	4735	4735	4735	4735	4735
Adjusted R^2	0.489	0.492	0.489	0.494	0.508
Mean of LEAP Assessment	-0.576	-0.576	-0.576	-0.576	-0.576
Standard Deviation of LEAP Assessment	1.010	1.010	1.010	1.010	1.010

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

In Models 3, 4 and 5, portfolio rankings of 2 are used as the comparison group, given that only 1-2 teachers were assigned a final portfolio ranking of 1 in each subject.

Average critical element scores are derived from the average of all critical element ratings given by all assessors and sources across the five framework areas. Teacher ranks are derived from the average score of all critical element ratings across all sources and evaluators for each teacher. Teachers are ranked from lowest to highest average critical element rating and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating of 1-4 by evaluators.

Teacher effect scores were determined by the above specification. Teachers were then ranked in order of lowest to highest teacher effect score, and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Similar to the way that Figure 1a graphically displayed the estimates from Panel A of Table 3 using the holistic PAS score (estimates reproduced in model 1 of Table 4a), Figure 2a (shown at end) gives a graphic representation of the estimates from model 2 of Table 4a. In Figure 2a teacher value-added scores in math are graphed against the constructed PAS scores used as the PAS measure in model 2 of Table 4a. Figures 2b-2e (shown at end) give the same graphical information for the other four subject areas.

The difference in the estimates between models 1 and 2 of Table 4a can occur through two routes. First, since the constructed PAS score takes on non-integer values, some of the within-category variation in effectiveness evident in figure 1 is utilized when the constructed measure is used. Second, it could be that scores at the critical element level are more closely related to teacher effectiveness than are holistically determined scores. To examine the extent to which the latter is true, we first use the constructed PAS scores to bin teachers into the same distributions as the end-of-year portfolio scores. To do this, we arrayed teachers according to their constructed cell average scores—least to greatest, and then, based on this array, we assigned teachers to have a "rank" of 1, 2, 3, or 4 so that the number of teachers within each score group matched the number generated by the PAS evaluators with their end-of-year scores. By doing this, we utilize the information content of the constructed PAS scores to rank teachers, but we throw away the additional variation generated by the non-integer values that result from the averaging process. This exercise addresses the following question: if the current distribution of teachers across categories—1, 2, 3, and 4—is the desirable distribution, how much more or less predictive is the current holistic scoring process relative to ranking teachers simply on the average of the critical element level cell entries.

A comparison of models 3 and 4 in Table 4a shows the relative performance of the constructed PAS to the holistically scored PAS in math when teachers in the math sample are, in both cases, constrained to the current distribution across the four categories. In each case—for categories 1, 3, and 4—the estimated relationship between a teacher's ranking and student achievement is stronger when teachers are binned based on the constructed PAS score (model 4) relative to the rankings assigned by PAS evaluators (model 3). This suggests that using the constructed PAS may be more informative than is the holistic PAS score. Another benchmark against which to judge the use of PAS is to ask how well do current methods compare to some "gold standard" PAS measure. To establish an ideal benchmark against which to judge the current implementation of the TNTP evaluation system with PAS, we order teachers least to greatest based on their value-added scores, so that the

teacher with the lowest value-added is first. And then, similar to what we did above, we assign teachers to PAS categories 1-4 to mimic the number of teachers who were assigned each PAS score in reality. As in model 4, indicators are then entered into equation 1 as teachers' PAS measures.

The results of this exercise for math are under model 5. We see that when teachers are ranked based on their ability to promote student achievement growth (i.e., their value-added), that teachers of rank 3 have, on average, students who scored 0.43 of a standard deviation higher on the end-of-year math tests than did the observationally similar students of teaches who were placed in the rank 2 category. Simply put, if PAS were (1) perfectly identifying teachers based on their ability to promote test score growth and (2) if we wanted to preserve the current distribution of teaches across bins, then we would predict that "good" teachers (those given a score of 3) generated more than four-tenths of a standard deviation in additional achievement gain relative to teachers PAS identified as "fair" teachers (those given a score of 2). In this exercise the current use of PAS predicts that the differential ability of a "good" versus a "fair" teacher in terms of promoting math achievement is less than half of what is predicted in the ideal (0.17 in model 3 versus 0.43 in model 5 on the category 3 indicator). Given this finding, the PAS evaluation system, as currently implemented, appears to be doing a poorer job at distinguishing teachers than is theoretically possible. Of course, we realize that use of the PAS has multiple objectives—only one of which is identifying effective teachers defined as those with a high value-added—and that there are logistical and political difficulties associated with changing PAS.

Yet another method for assessing the predictive value of PAS scores across specifications and source of the score is through the amount of between-teacher variation in student achievement that gets explained across the different models. For this exercise we first determine the total

between-teacher variation in conditional student achievement. We do this by estimating equation 1 without the PAS variable in a multi-level model to get a measure of the conditional variation that exists at the teacher level (level 2). We then estimate the same multilevel model including PAS as a predictor and then examine how much of the conditional between-teacher variance is explained by the addition of PAS to the model. For the math sample, PAS, as measured by the holistic assessment of evaluators, is entered linearly into equation 1 (i.e., the model 1 specification) and explains about 5%of the conditional between-teacher variation. By comparison, when the constructed PAS measure is used in the linear specification (model 2), PAS scores explain 16%of the between-teacher variation in student achievement. Thus, from this measure of predictability, the

constructed version of PAS explains substantially more of the variance in teacher effectiveness than does the evaluator-based measure of PAS. Even more pronounced results are found across models 3 and 4, with the PAS measure explaining 3%of the between-teacher in model 3 that uses the evaluator-based PAS measure and 24% in model 4 that uses the constructed PAS measure.

Turning now to the other four subjects, Tables 4b-4e present estimates that allow for the same set of comparisons as was just discussed for math. A careful study of the results across the different subjects suggests two main conclusions.

TABLE 4B. COMPARISONS OF THE ELA ACHIEVEMENT-PAS RELATIONSHIP ACROSS MODELS

	Model 1	Model 2	Model 3	Model 4	Model 5
Final Portfolio Rating	0.0421 (0.0402)				
Average of All Critical Element Scores		0.121 (0.0639)			
Final Portfolio Rating of 1			0.174 (0.106)		
Final Portfolio Rating of 3			0.0581 (0.0457)		
Final Portfolio Rating of 4			0.314** (0.109)		
Average Critical Element Score Rank of 1				-0.304 (0.259)	
Average Critical Element Score Rank of 3				0.0725 (0.0427)	
Average Critical Element Score Rank of 4				0.336** (0.105)	
Teacher Effect Rank of 1					-0.324*** (0.0672)
Teacher Effect Rank of 3					0.384*** (0.0365)
Teacher Effect Rank of 4					1.092*** (0.173)

TABLE 4B. CONTINUED

Teacher Observations	107	107	107	107	107
Student Observations	4040	4040	4040	4040	4040
Adjusted R^2	0.533	0.533	0.533	0.534	0.548
Mean of LEAP Assessment	-0.582	-0.582	-0.582	-0.582	-0.582
Standard Deviation of LEAP Assessment	1.033	1.033	1.033	1.033	1.033

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

In Models 3, 4 and 5, portfolio rankings of 2 are used as the comparison group, given that only 1-2 teachers were assigned a final portfolio ranking of 1 in each subject.

Average critical element scores are derived from the average of all critical element ratings given by all assessors and sources across the five framework areas. Teacher ranks are derived from the average score of all critical element ratings across all sources and evaluators for each teacher. Teachers are ranked from lowest to highest average critical element rating and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating of 1-4 by evaluators.

Teacher effect scores were determined by the above specification. Teachers were then ranked in order of lowest to highest teacher effect score, and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 4C. COMPARISONS OF THE ELA ACHIEVEMENT-PAS RELATIONSHIP ACROSS MODELS

	Model 1	Model 2	Model 3	Model 4	Model 5
Final Portfolio Rating	0.143* (0.0569)				
Average of All Critical Element Scores		0.297*** (0.0558)			
Final Portfolio Rating of 1			-0.147 (0.283)		
Final Portfolio Rating of 3			0.132* (0.0522)		
Final Portfolio Rating of 4			0.526*** (0.122)		
Average Critical Element Score Rank of 1				-0.619*** (0.145)	
Average Critical Element Score Rank of 3				0.0574 (0.0487)	
Average Critical Element Score Rank of 4				0.127 (0.112)	
Teacher Effect Rank of 1					-0.262** (0.0782)
Teacher Effect Rank of 3					0.394*** (0.0493)
Teacher Effect Rank of 4					1.061*** (0.126)
Teacher Observations	64	64	64	64	64
Student Observations	2271	2271	2271	2271	2271
Adjusted R^2	0.413	0.416	0.412	0.416	0.425
Mean of LEAP Assessment	-0.648	-0.648	-0.648	-0.648	-0.648
Standard Deviation of LEAP Assessment	1.071	1.071	1.071	1.071	1.071

TABLE 4C. CONTINUED

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

In Models 3, 4 and 5, portfolio rankings of 2 are used as the comparison group, given that only 1-2 teachers were assigned a final portfolio ranking of 1 in each subject.

Average critical element scores are derived from the average of all critical element ratings given by all assessors and sources across the five framework areas. Teacher ranks are derived from the average score of all critical element ratings across all sources and evaluators for each teacher. Teachers are ranked from lowest to highest average critical element rating and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating of 1-4 by evaluators.

Teacher effect scores were determined by the above specification. Teachers were then ranked in order of lowest to highest teacher effect score, and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 4D. COMPARISONS OF THE SCIENCE ACHIEVEMENT-PAS RELATIONSHIP ACROSS MODELS

	Model 1	Model 2	Model 3	Model 4	Model 5
Final Portfolio Rating	0.161*** (0.0443)				
Average of All Critical Element Scores		0.274** (0.0832)			
Final Portfolio Rating of 1			0.209 (0.133)		
Final Portfolio Rating of 3			0.184*** (0.0467)		
Final Portfolio Rating of 4			0.430 (0.303)		
Average Critical Element Score Rank of 1				0.134 (0.130)	
Average Critical Element Score Rank of 3				0.180*** (0.0490)	
Average Critical Element Score Rank of 4				0.439 (0.288)	
Teacher Effect Rank of 1					-0.317*** (0.0889)
Teacher Effect Rank of 3					0.506*** (0.0376)
Teacher Effect Rank of 4					1.017*** (0.0899)
Teacher Observations	83	83	83	83	83
Student Observations	4070	4070	4070	4055	4070
Adjusted R^2	0.422	0.422	0.422	0.422	0.440

TABLE 4D. CONTINUED

Mean of LEAP Assessment	-0.499	-0.499	-0.499	-0.499	-0.499
Standard Deviation of LEAP Assessment	0.997	0.997	0.997	0.997	0.997

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

In Models 3, 4 and 5, portfolio rankings of 2 are used as the comparison group, given that only 1-2 teachers were assigned a final portfolio ranking of 1 in each subject.

Average critical element scores are derived from the average of all critical element ratings given by all assessors and sources across the five framework areas. Teacher ranks are derived from the average score of all critical element ratings across all sources and evaluators for each teacher. Teachers are ranked from lowest to highest average critical element rating and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating of 1-4 by evaluators.

Teacher effect scores were determined by the above specification. Teachers were then ranked in order of lowest to highest teacher effect score, and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 4E. COMPARISONS OF THE SOCIAL STUDIES ACHIEVEMENT-PAS RELATIONSHIP ACROSS MODELS

	Model 1	Model 2	Model 3	Model 4	Model 5
Final Portfolio Rating	0.247** (0.0864)				
Average of All Critical Element Scores		0.284* (0.122)			
Final Portfolio Rating of 1			-0.479* (0.223)		
Final Portfolio Rating of 3			0.130 (0.116)		
Final Portfolio Rating of 4			0.920*** (0.213)		
Average Critical Element Score Rank of 1				-0.351 (0.202)	
Average Critical Element Score Rank of 3				0.283*** (0.0816)	
Average Critical Element Score Rank of 4				1.113*** (0.189)	
Teacher Effect Rank of 1					-0.261* (0.121)
Teacher Effect Rank of 3					0.475*** (0.0482)
Teacher Effect Rank of 4					1.152*** (0.128)
Teacher Observations	59	59	59	59	59
Student Observations	2146	2146	2146	2146	2146
Adjusted R^2	0.361	0.359	0.362	0.366	0.382
Mean of LEAP Assessment	-0.509	-0.509	-0.509	-0.509	-0.509
Standard Deviation of LEAP Assessment	1.052	1.052	1.052	1.052	1.052

TABLE 4E. CONTINUED

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

In Models 3, 4 and 5, portfolio rankings of 2 are used as the comparison group, given that only 1-2 teachers were assigned a final portfolio ranking of 1 in each subject.

Average critical element scores are derived from the average of all critical element ratings given by all assessors and sources across the five framework areas. Teacher ranks are derived from the average score of all critical element ratings across all sources and evaluators for each teacher. Teachers are ranked from lowest to highest average critical element rating and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating of 1-4 by evaluators.

Teacher effect scores were determined by the above specification. Teachers were then ranked in order of lowest to highest teacher effect score, and subsequently assigned a rank of 1-4. The number of teachers assigned to each category is equal to the number of teachers who received the same final portfolio rating.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

First, there is evidence that in math and social studies the constructed measures are better predictors of teacher effectiveness than are the holistic assessments of the evaluators. In these subjects, the estimates on the category 3 indicator are roughly twice as large in model 4 that uses the constructed measure than in model 3 that uses the evaluator assessment as the PAS measure. At the same time, this pattern is not evident in the other three subjects.

A second lesson to draw from Tables 4a-4e is that for every subject the current PAS system falls short of what could be done in the ideal in terms of identifying effective teachers. This is seen by comparing the estimates on the category 3 indicator between models 3 and 5.

B. PAS Framework Scores as Measures of Teacher Effectiveness

RESULTS FROM THE SECTION above point out that how one aggregates information across the critical elements can affect, and perhaps improve, how predictive PAS scores are. Given that the critical element scores reside within frameworks, this suggests that certain frameworks may be more predictive of teacher effectiveness than others.

To examine this question we estimate equation 1 using the PAS vector scores from each of the five PAS frameworks. To begin with, our measure of framework scores is based on the framework scores assigned by the evaluators at the end of the year. When there was disagreement, we used the average of the evaluators' framework scores. Table 5a presents estimates for each subject from this construction of framework scores.

TABLE 5A. ESTIMATED RELATIONSHIP BETWEEN STUDENT ACHIEVEMENT GROWTH AND FRAMEWORK RATINGS BASED ON THE AVERAGE OF FINAL FRAMEWORK AREA SCORES ASSIGNED BY ALL EVALUATORS

	Math	ELA	Reading	Science	Social Studies
Class Environment	0.130* (0.0501)	0.105 (0.0538)	0.177* (0.0705)	0.0122 (0.0543)	0.0626 (0.0728)
Instructional Design and Delivery	0.225** (0.0767)	0.165* (0.0642)	-0.0555 (0.112)	0.0464 (0.106)	-0.0577 (0.119)
Assessment	0.00450 (0.0632)	-0.0769 (0.0616)	0.0634 (0.105)	0.0392 (0.0465)	-0.304* (0.115)
Professionalism	-0.0979 (0.0929)	-0.0910 (0.0618)	-0.0493 (0.0661)	0.0767 (0.0586)	0.134 (0.116)
Student Achievement	-0.0398 (0.0603)	-0.0331 (0.0514)	0.0953 (0.0752)	0.0155 (0.0701)	0.268** (0.100)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146
Adjusted R^2	0.493	0.535	0.416	0.421	0.362
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimates in the first column indicate that scores on Class Environment (CE) and Instructional Design and Delivery (IDD) are the only significant predictors of teachers' effectiveness in promoting student achievement growth in math. A one-unit increase in a teacher's final Class Environment score is associated with 0.13 of a standard deviation greater achievement growth in math, while the effect size for IDD is 0.23. The estimated coefficients on the other three frameworks in math are smaller and none are statistically significant.

Class Environment is the only statistically significant predictor in ELA, and IDD is the only significant predictor in reading; these are the only subjects where scores on these two frameworks are predictive of student achievement growth. There are no framework scores that are predictive of teacher effectiveness in science. Meanwhile, there are competing stories in social studies. On the one hand teachers who score higher on the student achievement framework are associated with higher than predicted student achievement scores in social studies, while on the other hand high scores on the assessment framework are associated with lower than predicted student achievement scores.

We next construct framework scores based on the within-teacher average of the critical element scores within each framework. As was the case earlier, we use this construction to take advantage of some of the within-category variation in teacher effectiveness evident in Figures 1a-1e. Estimates based on this construction of framework scores are presented in Table 5b.

TABLE 5B. ESTIMATED RELATIONSHIP BETWEEN STUDENT ACHIEVEMENT GROWTH AND FRAMEWORK RATINGS BASED ON THE AVERAGE OF ALL CRITICAL ELEMENT RATINGS

	Math	ELA	Reading	Science	Social Studies
Class Environment	0.0855 (0.0827)	0.122 (0.0817)	0.163 (0.0989)	0.0472 (0.0987)	0.181 (0.125)
Instructional Design and Delivery	0.456** (0.154)	0.221 (0.126)	0.160 (0.157)	-0.134 (0.197)	-0.136 (0.320)
Assessment	0.0847 (0.128)	-0.0491 (0.102)	-0.0161 (0.118)	0.140 (0.0888)	-0.343 (0.280)
Professionalism	-0.256* (0.106)	-0.116 (0.0871)	-0.107 (0.128)	0.246** (0.0746)	0.403 (0.223)
Student Achievement	0.00911 (0.0755)	-0.0871 (0.0770)	0.0667 (0.0830)	-0.0330 (0.0690)	0.0700 (0.183)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146
Adjusted R^2	0.494	0.535	0.416	0.422	0.361
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

Standard errors in parentheses
Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.
Average framework scores represent the mean of all critical element scores within each framework area assigned by all evaluators across all sources for an individual teacher. Evaluators assigned a rating of 1-4 for each of the five framework areas.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The only robust finding across Tables 5a and 5b is that higher IDD scores are predictive of a teacher's ability to generate higher student achievement in math. The only other statistically significant relationships in Table 5b are with the Professionalism framework. Based on the constructed framework scores in Table 5b, higher scores on the Professionalism framework are associated with lower than predicted student achievement grown in math, but higher achievement growth in science. We have no good explanation of these different findings within this framework and across subjects, and, given that we are comparing many different estimates, some could be statistically significant by chance. Nevertheless, there is some evidence across Tables 5a and 5b that high scores in some framework areas may be predictive of lower teacher effectiveness. These potential negative relationships deserve careful thought and more study.

Given the importance of IDD for math, a relevant question is whether there are particular elements that are driving the positive IDD correlations in math. To examine this question we replace the constructed IDD framework score with the within-teacher, across-source average of each critical element within the IDD framework. Estimates from this composition of the PAS vector are presented in Table 5c. Based on the results in the first column it appears that the Structure of Lesson and Student Engagement critical elements are the primary drivers behind the predictive nature of IDD for teacher effectiveness in math.

TABLE 5C. ESTIMATED RELATIONSHIPS BETWEEN STUDENT ACHIEVEMENT GROWTH AND FRAMEWORK RATINGS BASED ON THE AVERAGE OF CRITICAL ELEMENT RATINGS WITHIN FRAMEWORK AREA AND AVERAGE RATINGS OF EACH CRITICAL ELEMENT WITHIN THE INSTRUCTIONAL DESIGN AND DELIVERY FRAMEWORK

	Math	ELA	Reading	Science	Social Studies
Class Environment	0.115 (0.0886)	0.118 (0.0835)	0.246* (0.105)	0.0173 (0.0910)	0.171 (0.134)
Instructional Design & Delivery					
Backwards Planning	0.0438 (0.0917)	0.113 (0.0853)	0.0486 (0.0982)	0.283** (0.106)	0.0385 (0.149)
General Student Instruction	-0.0621 (0.0966)	-0.0368 (0.114)	0.0696 (0.157)	-0.130 (0.101)	0.0542 (0.210)
Modify Instruction	0.0554 (0.102)	0.0123 (0.102)	0.0862 (0.0962)	-0.301** (0.114)	-0.370 (0.193)
Structure of Lesson	0.190** (0.0638)	0.0619 (0.0890)	0.0499 (0.0992)	0.152 (0.0942)	-0.0490 (0.196)
Student Engagement	0.194** (0.0695)	0.0612 (0.0770)	-0.209* (0.103)	0.00176 (0.114)	0.186 (0.248)
Assessment	0.145 (0.127)	-0.0434 (0.123)	-0.0401 (0.133)	0.149 (0.0857)	-0.342 (0.264)
Professionalism	-0.292** (0.108)	-0.100 (0.0891)	-0.0142 (0.130)	0.177 (0.0910)	0.283 (0.251)
Student Achievement	-0.00948 (0.0735)	-0.0825 (0.0774)	0.0424 (0.100)	-0.0186 (0.0828)	0.154 (0.195)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146
Adjusted R^2	0.495	0.534	0.416	0.426	0.362
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.
Average framework scores represent the mean of all critical elements scores across all sources and evaluators within each framework area for an individual teacher. The Instructional Design and Delivery framework is divided into each of its critical elements which represent the average of critical element scores across all sources and evaluators for each teacher.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We go through a similar exercise to try to better understand the negative relationship between Professionalism and math achievement on the one hand and the positive relationship between Professionalism and social studies achievement on the other. In the first column of Table 5d we see that each of the three critical elements in the Professionalism framework are negatively related to math achievement, though only Fulfills Responsibilities is statistically significant. Meanwhile, none of the three critical elements that comprise Professionalism stand out as driving the positive relationship between Professionalism framework scores and student achievement gains. It does appear that Positive Relationships is an important critical element for predicting social studies achievement.

TABLE 5D. ESTIMATED RELATIONSHIPS BETWEEN STUDENT ACHIEVEMENT GROWTH AND FRAMEWORK RATINGS BASED ON THE AVERAGE OF CRITICAL ELEMENT RATINGS WITHIN FRAMEWORK AREA AND AVERAGE RATINGS OF EACH CRITICAL ELEMENT WITHIN THE PROFESSIONALISM FRAMEWORK

	Math	ELA	Reading	Science	Social Studies
Class Environment	0.0606 (0.0817)	0.140 (0.0912)	0.171 (0.109)	0.0595 (0.111)	-0.0342 (0.138)
Instructional Design & Delivery	0.393* (0.157)	0.224 (0.125)	0.162 (0.159)	-0.114 (0.198)	-0.0247 (0.296)
Assessment	0.0933 (0.128)	-0.0697 (0.106)	-0.0184 (0.126)	0.128 (0.0930)	-0.369 (0.270)
Professionalism					
Reflects on Practice	-0.0620 (0.0602)	-0.000493 (0.0628)	-0.0288 (0.0904)	0.101 (0.0617)	-0.0470 (0.115)
Positive Relationships	-0.0397 (0.0844)	-0.0994 (0.0894)	-0.0575 (0.102)	0.0557 (0.0846)	0.560** (0.182)
Fulfills Responsibilities	-0.128* (0.0526)	-0.0326 (0.0317)	-0.0268 (0.106)	0.0700 (0.0401)	0.0244 (0.119)
Student Achievement	0.0329 (0.0770)	-0.0886 (0.0752)	0.0620 (0.105)	-0.0155 (0.0746)	0.0105 (0.181)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146
Adjusted R^2	0.494	0.534	0.415	0.422	0.364
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

Standard errors in parentheses

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.

Average framework area scores represent the mean of all critical elements across all sources and evaluators within each framework area for each teacher. The Professionalism framework is divided into each of its critical elements which are represented by the average critical element score across all sources and evaluators for each teacher.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

C. Sources of Evidence as Measures of Teacher Effectiveness

AS WAS THE CASE for critical elements, we are interested in whether there are particular sources of evidence that are more predictive of teacher effectiveness than others. To construct measures of evidence source we use the within-teacher, within-source average of the scores evaluators assigned at the critical element level. Table 6a provides estimates of equation 1 where the PAS vector is composed of a teacher's average critical element score within each evidence source: Instructional Unit, Video, Observation, DSA Report,[27] and Survey.

TABLE 6A. ESTIMATED RELATIONSHIPS BETWEEN STUDENT ACHIEVEMENT GROWTH AND THE AVERAGE OF CRITICAL ELEMENT RATINGS BY SOURCE

	Math	ELA	Reading	Science	Social Studies
Instructional Unit	0.312** (0.103)	-0.0517 (0.0913)	0.0143 (0.0676)	0.109 (0.100)	-0.249 (0.225)
Video	0.0296 (0.0760)	0.0395 (0.0453)	0.0894 (0.0643)	0.0735 (0.0380)	0.388* (0.148)
Observation	0.188* (0.0833)	0.184* (0.0780)	0.0565 (0.0915)	-0.0827 (0.0580)	-0.120 (0.181)
DSA Report	-0.0846 (0.0665)	-0.0701 (0.0447)	0.0207 (0.0526)	0.103* (0.0456)	0.0875 (0.164)
Survey	-0.152 (0.0775)	0.0173 (0.0939)	0.110 (0.0932)	0.0120 (0.0616)	0.0626 (0.144)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146
Adjusted R^2	0.494	0.534	0.415	0.423	0.362
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

[27] In the DSA report teachers describe and document evidence of improved academic achievement of their students. Teachers are free to set their own goals as well as to determine and administer diagnostics of their choosing to measure achievement of those goals. These reports do not include state standardized test scores and are not standardized objective measures of student achievement. Rather, achievement reports are based on judgments made by each individual teacher about the progress of his or her students as measured by formative assessments that either the teacher develops or adapts from a teaching resource.

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.
Source scores represent an average of all critical element scores recorded from a particular source across all evaluators for a particular teacher. The CSL form is not included as a source in this analysis.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results for math in the first column of Table 6a indicate that teachers who score higher on critical elements within both the Instructional Unit and the Observation sources tend to produce greater student achievement growth, about three-tenths of a standard deviation more in the case of Instructional Unit. We note that the coefficient on Survey is also negative and of approximately the same size as the positive estimate on Observation though not quite statistically significant.

As is the case in math, the Observation source is positively related to teacher effectiveness in ELA. The only other instances where any of the evidence sources are statistically significant are in science and social studies. In science the DSA Report evidence source is positively related to student achievement growth, while in social studies higher scores on the Video source of evidence are related to greater achievement growth.

Finally, in Table 6b we examine the extent to which the predictive quality of the source of evidence might be particular to a given framework area. In that table the elements of the PAS vector are the within-teacher, within-source, within-framework averages of the critical element scores.

TABLE 6B. ESTIMATED RELATIONSHIPS BETWEEN STUDENT ACHIEVEMENT GROWTH AND THE AVERAGE OF ALL CRITICAL ELEMENT RATINGS WITHIN SOURCE AND FRAMEWORK AREA

	Math	ELA	Reading	Science	Social Studies
Classroom Environment					
Observation	-0.0508 (0.0442)	0.0399 (0.0659)	0.0189 (0.0710)	-0.0259 (0.0639)	0.0258 (0.106)
Survey	0.0285 (0.0634)	0.0189 (0.0658)	-0.116 (0.0999)	-0.00439 (0.0929)	-0.397* (0.196)
Video	0.136* (0.0599)	-0.0358 (0.0875)	0.210 (0.122)	0.192* (0.0834)	0.123 (0.173)
Instructional Design & Delivery					
Instructional Unit	0.379*** (0.0781)	0.0372 (0.0770)	0.202 (0.136)	0.0320 (0.121)	-0.0988 (0.186)
Observation	0.0341 (0.0709)	-0.128 (0.0714)	-0.00701 (0.0979)	-0.0314 (0.0726)	-0.117 (0.160)
Survey	-0.170* (0.0832)	0.128 (0.103)	0.0137 (0.166)	0.104 (0.0861)	0.274 (0.197)
Video	0.00727 (0.0825)	0.120 (0.0944)	0.0358 (0.131)	0.0782 (0.0914)	0.287 (0.195)
Assessment					
DSA Report	-0.0598 (0.0613)	-0.0410 (0.0704)	0.0218 (0.0815)	0.0909 (0.0553)	-0.154 (0.173)
Instructional Unit	0.0615 (0.0774)	-0.0392 (0.0574)	-0.0457 (0.0916)	0.153 (0.0801)	-0.289 (0.149)
Observation	0.216*** (0.0502)	0.191** (0.0637)	0.0790 (0.0685)	-0.119 (0.0773)	-0.0575 (0.120)
Survey	-0.0174 (0.0582)	-0.0521 (0.0909)	0.00982 (0.144)	-0.141 (0.131)	0.248 (0.207)
Professionalism					
Observation	0.00541 (0.0423)	-0.00180 (0.0454)	-0.111 (0.0873)	0.00235 (0.0543)	-0.0239 (0.147)
Survey	-0.0955 (0.0671)	-0.0767 (0.0759)	0.111 (0.103)	0.0299 (0.0865)	-0.0123 (0.136)
Video	-0.106 (0.0678)	-0.0698 (0.0660)	-0.102 (0.125)	0.0356 (0.0857)	0.172 (0.169)
Student Achievement					
DSA Report	0.0454 (0.0564)	-0.0552 (0.0600)	0.0377 (0.0637)	0.152* (0.0614)	0.393** (0.131)
Instructional Unit	0.00754 (0.0600)	0.0720 (0.0507)	-0.0137 (0.0788)	-0.0183 (0.0497)	-0.120 (0.106)
Survey	0.0600 (0.0500)	-0.0539 (0.0498)	0.0785 (0.0743)	0.0730 (0.0621)	0.0492 (0.0848)
Video	-0.119* (0.0571)	0.0256 (0.0515)	-0.0913 (0.0645)	-0.249** (0.0755)	-0.126 (0.146)
Teacher Observations	109	107	64	83	59
Student Observations	4735	4040	2271	4070	2146

TABLE 6B. CONTINUED

Adjusted R^2	0.499	0.537	0.415	0.427	0.368
Mean of LEAP Assessment	-0.576	-0.582	-0.648	-0.499	-0.509
Standard Deviation of LEAP Assessment	1.010	1.033	1.071	0.997	1.052

Note: Each column represents a separate student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Clustered (teacher) standard errors in parentheses.
Source scores within each framework area represent an average of critical element scores across all evaluators within each framework area and source for an individual teacher. The CSL form is not included as a source in this analysis.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Looking down the first column it appears that Instructional Unit as a positive predictor of teacher effectiveness in math is only important within the IDD framework. Meanwhile, higher scores on the Video evidence source are predictive of higher math achievement within Classroom Environment and lower scores within the Student Achievement framework. Still looking down the first column, Observation as a source of evidence predicts positively within the Assessment framework, while the Survey evidence source predicts negatively within IDD.

Looking at the within-framework evidence source estimates across the other subject areas shows positive and negative point estimates, only a few of which are statistically significant; no clear patterns emerge from the table. The estimates in Table 6b are meant to be suggestive and to provoke further thought. We caution against making too much of these results. Given that we present 90-point estimates in the table that are not independent, we would expect a number of estimates to be significant by chance.

VI. Discussion and Conclusion

IN THE PRECEDING ANALYSIS, we have examined PAS scores of Louisiana teachers from 2005 to 2009, and the relationship between teacher PAS scores and student achievement growth.

We find a modest, positive correlation between teacher PAS scores and student achievement growth in math and reading. There is not strong evidence of any relationship between PAS scores and achievement in the other subjects, although our estimates are not precise enough to rule out moderately large effects. The math and reading results are generally consistent with prior studies that compare classroom observation protocols and student achievement growth, including the recently released findings from the Measures of Effective Teaching (MET) project.

Interestingly, we find that there is quite little variation in PAS scores across teachers. That is, the vast majority of teachers receive a score of 2 or 3 on the PAS.[28] The limited variation in PAS scores appears to be one reason that the relationship between PAS scores and achievement are not higher. When we create a PAS score by averaging all critical element scores, this "constructed PAS" measure has considerably more variation since it is not limited to integer values and is more strongly associated with student achievement than the original (integer-value) assessor-based PAS score.

These results, along with our review of the PAS components, suggest several paths that TNTP might pursue in order to strengthen the relationship between teacher evaluation scores and student achievement growth. Perhaps most importantly, TNTP could experiment with ways to create greater variation in PAS scores. This might be accomplished in a number of ways, including expanding the set of rating categories, providing raters a rough "curve" to follow, or providing raters with more specific guidance with regard to what constitutes excellence.

In addition, TNTP could change how PAS scores are determined, either by adding/subtracting specific critical elements or data sources, or by changing how critical element scores are combined to produce a final rating. For example, we find that there are differences between the constructed and assessor–based PAS measures that arise from how assessors weigh the critical elements coming up with an overall score. In math, ELA, and

[28] The Measures of Effective Teaching Project has found greater variation in scores with several classroom observation tools using trained outside observers; See Figure 8, page 26-27, of Kane, Thomas and Doug Staiger. Gathering Feedback for Teaching (Bill and Melinda Gates Foundation, 2012). The Widget Effect, however, found little variation in actual teacher evaluation scores; See Weisberg, Daniel, Sexton, Susan, Mulhern, Jennifer, Keeling, David. THE WIDGET EFFECT: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness (The New Teacher Project, 2009).

science, the four-category assessor-based PAS score is less predictive of student achievement growth than an analogous four-category constructed PAS measure, although this pattern is not evident in reading or social studies. Similarly, we see that Instructional Design and Delivery shows the strongest relationship to student achievement in math and (to a lesser extent) science. For ELA and reading, the class environment score shows the largest positive relationship to student achievement growth. We recommend that TNTP further explore these and related issues in an effort to redesign the PAS in a way that maximizes its power to predict student achievement growth.

Another important finding involves inter-rater reliability (IRR). We find that the IRR of the PAS framework scores and critical element scores is considerably lower than what is typically recommended in the psychometrics literature. It is likely that such low reliabilities are weakening the explanatory power of the PAS. We suggest TNTP explore ways to increase the inter-rater reliability in the future by, for example, deepening the training of raters.

Finally, we note that these results have implications for the large number of states, districts, and non profits currently redesigning their teacher evaluation systems. The findings from this and other work the authors have conducted suggest the importance that seemingly innocuous design decisions can have on evaluation systems. For instance, allowing raters to combine their subscores into "holistic" domain/framework ratings may cause some of the information and implicit rankings captured by specific elements to be lost. This could be due to there being less "wiggle room" at the subscore level (which requires very specific types of evidence) than at the domain/framework level where irrelevant favorable or unfavorable impressions of those being evaluated may play an unconscious role in scoring. Designers may thus want to consider providing less freedom around how individual elements are combined into summary scores to raters. Similarly, forcing teacher ratings into only a few buckets necessarily gives up much of the information painstakingly gathered about relative teacher performance and may hide specific tasks or competencies on which groups of teachers in a system are struggling. Designers of teacher evaluation systems should consider strategies to take maximum advantage of the detailed information they are requiring evaluators to collect both in order to better understand the range of teacher performance and to better guide teacher professional development and growth.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95-135.

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011) An interaction-based approach to enhancing secondary school instruction and student achievement. *Science* 333(6045):1034–37.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying Effective Teachers Using Performance on the Job. Washington, DC: The Brookings Institution.

Holtzapple, Elizabeth. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education,* 17(3): 207-219.

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., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources*, *46*(3), 587-613.

Kane, Thomas and Doug Staiger. Gathering Feedback for Teaching. (Bill and Melinda Gates Foundation, 2012).

Landis, J.R.; & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1): 159–174.

Milanowski, Anthony. (2004a). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*. Vol. 79, No. 4, pp. 33-53.

Milanowski, Anthony. (2004b). Relationships among dimension scores of standards-based teacher evaluation systems, and the stability of evaluation score—student achievement relationships over time. *Consortium for Policy Research in Education- University of Wisconsin Working Paper Series*, No. TC-04-02, April 2004.

Rivkin, S., Hanushek, R., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

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

Sanders, W.L., & Rivers, J.C. (1996). Cumulative and Residual Effects of Teachers on Future Student Academic Achievement. University of Tennessee Value-Added Research and Assessment Center. November.

Weisberg, Daniel, Sexton, Susan, Mulhern, Jennifer, Keeling, David. THE WIDGET EFFECT: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. (The New Teacher Project, 2009).

Additional Tables and Figures

TABLE 2B, PANEL B. CORRELATION MATRIX OF CLASSROOM ENVIRONMENT CRITICAL ELEMENTS AS GIVEN BY EVALUATORS

	Systems & Routines: Video	Systems & Routines: Observation	Communicates Standards: Video	Communicates Standards: Surveys	Communicates Standards: Observation	Physical Environment: Video	Positive Environment: Video	Positive Environment: Survey	Positive Environment: Observation
Systems & Routines - Video	1								
Systems & Routines - Observation	0.422	1							
Communicates Standards - Video	0.824	0.428	1						
Communicates Standards - Survey	0.245	0.198	0.296	1					
Communicates Standards - Observation	0.388	0.533	0.415	0.247	1				
Physical Environment - Video	0.634	0.328	0.613	0.266	0.321	1			
Positive Environment - Video	0.777	0.454	0.809	0.287	0.393	0.672	1		
Positive Environment - Survey	0.384	0.297	0.382	0.626	0.275	0.380	0.442	1	
Positive Environment - Observation	0.357	0.467	0.373	0.267	0.570	0.355	0.361	0.325	1

TABLE 2B, PANEL C. CORRELATION MATRIX OF INSTRUCTIONAL DESIGN & DELIVERY CRITICAL ELEMENTS AS GIVEN BY EVALUATORS

Backwards Plan: Instructional Unit	1																					
Backwards Plan: Survey	0.148	1																				
Backwards Plan: Observation	0.291	0.160	1																			
General Student Instruction: Instructional Unit	0.635	0.119	0.436	1																		
General Student Instruction: Video	0.303	0.315	0.345	0.379	1																	
General Student Instruction: Survey	0.272	0.229	0.170	0.297	0.203	1																
General Student Instruction: Observation	0.343	0.107	0.546	0.474	0.328	0.342	1															
Modifies Instruction: Instructional Unit	0.475	0.146	0.316	0.620	0.383	0.193	0.332	1														
Modifies Instruction: Video	0.268	0.276	0.309	0.412	0.825	0.249	0.292	0.375	1													
Modifies Instruction: Survey	0.273	0.562	0.210	0.265	0.422	0.373	0.146	0.312	0.387	1												
Modifies Instruction: Observation	0.215	0.139	0.388	0.338	0.267	0.212	0.523	0.343	0.273	0.212	1											
Structure Lesson: Instructional Unit	0.606	0.114	0.377	0.731	0.380	0.359	0.442	0.561	0.398	0.300	0.283	1										
Structure Lesson: Video	0.261	0.238	0.368	0.450	0.699	0.223	0.405	0.404	0.698	0.349	0.290	0.468	1									
Student Engagement: Instructional Unit	0.376	0.0730	0.331	0.530	0.466	0.174	0.367	0.575	0.439	0.302	0.225	0.514	0.369	1								
Student Engagement: Video	0.208	0.327	0.286	0.371	0.875	0.251	0.289	0.372	0.851	0.394	0.256	0.325	0.722	0.465	1							
Student Engagement: Survey	0.211	0.474	0.222	0.201	0.365	0.220	0.0692	0.246	0.392	0.623	0.176	0.143	0.292	0.162	0.360	1						

TABLE 2B, PANEL D. CORRELATION MATRIX OF ASSESSMENT CRITICAL ELEMENTS AS GIVEN BY EVALUATORS

	Selects Assessments: Instructional Unit	Selects Assessments: DSA	Selects Assessments: Surveys	Selects Assessments: Observation	Monitors Progress: Instructional Unit	Monitors Progress: DSA	Modifies Instruction: Instructional Unit	Modifies Instruction: DSA	Modifies Instruction: Surveys	Communicates Progress: Instructional Unit	(mean) assess_cp_survey	(mean) assess_cp_obs
Selects Assessments: Instructional Unit	1											
Selects Assessments: DSA	0.545	1										
Selects Assessments: Survey	0.196	0.160	1									
Selects Assessments: Observation	0.373	0.349	0.243	1								
Monitors Progress: Instructional Unit	0.529	0.258	0.197	0.330	1							
Monitors Progress: DSA	0.429	0.584	0.0913	0.234	0.362	1						
Modifies Instruction: Instructional Unit	0.471	0.371	0.190	0.227	0.377	0.408	1					
Modifies Instruction: DSA	0.356	0.616	0.117	0.366	0.298	0.621	0.394	1				
Modifies Instruction: Survey	0.287	0.181	0.210	0.182	0.202	0.202	0.131	0.149	1			
Communicates Progress: Instructional Unit	0.586	0.416	0.215	0.294	0.430	0.408	0.445	0.382	0.0527	1		
Communicates Progress: Survey	0.291	0.216	0.702	0.215	0.197	0.190	0.238	0.237	0.263	0.279	1	
Communicates Progress: Observation	0.345	0.266	0.270	0.395	0.344	0.251	0.258	0.293	0.288	0.269	0.226	1

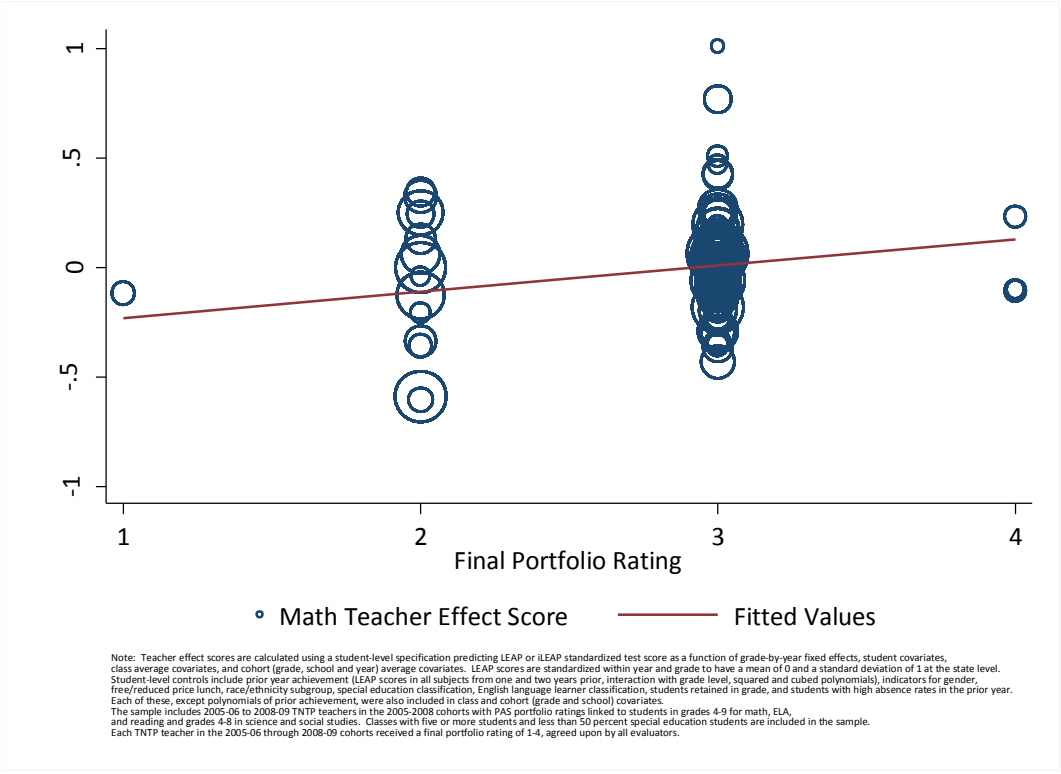
TABLE 2B, PANEL E. CORRELATION MATRIX OF PROFESSIONALISM CRITICAL ELEMENTS AS GIVEN BY EVALUATORS

	Reflects on Practice: Video	Reflects on Practice: Observation	Positive Relationships: Video	Positive Relationships: Surveys	Fulfills Responsibilities: CSL Form
Reflects on Practice: Video	1				
Reflects on Practice: Observation	0.391	1			
Positive Relationships: Video	0.628	0.299	1		
Positive Relationships: Survey	0.311	0.333	0.309	1	
Fulfills Responsibilities: CSL Form	0.269	0.224	0.215	0.202	1

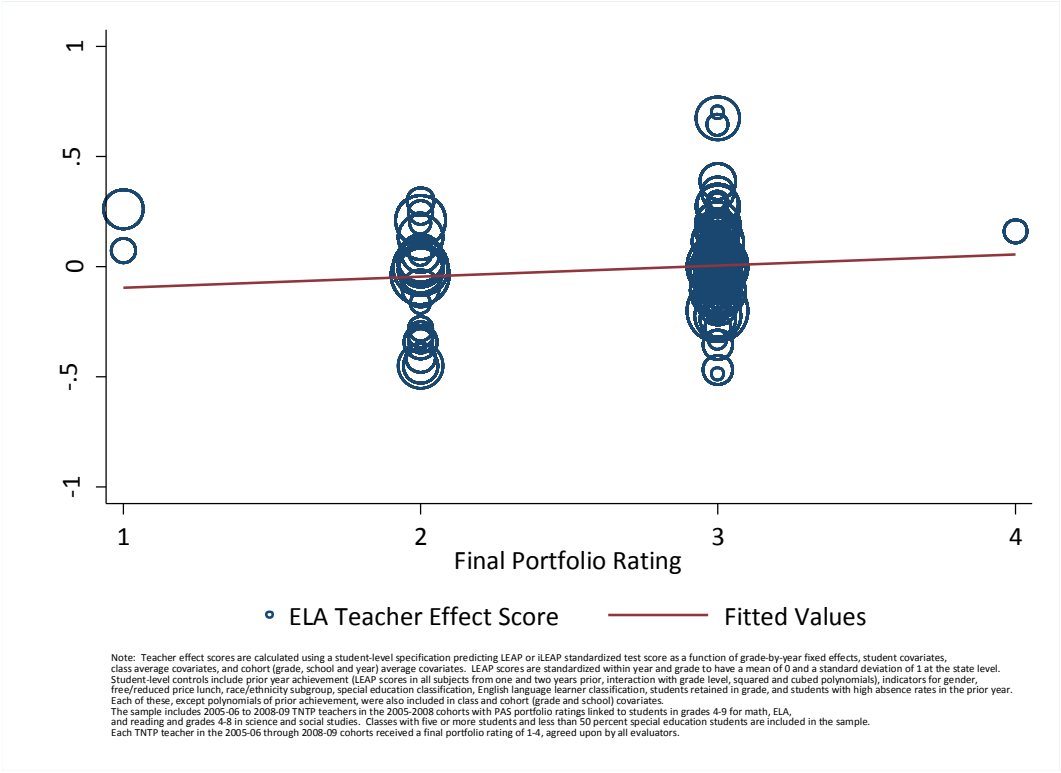
TABLE 2B, PANEL F. CORRELATION MATRIX OF STUDENT ACHIEVEMENT CRITICAL ELEMENTS AS GIVEN BY EVALUATORS

	Evidence Progress: Instructional Unit	Evidence Progress: Video	Evidence Progress: DSA	Evidence Progress: Surveys
Evidence Progress: Instructional Unit	1			
Evidence Progress: Video	0.481	1		
Evidence Progress: DSA	0.635	0.452	1	
Evidence Progress: Survey	0.182	0.193	0.184	1

FIGURE 1A. RELATIONSHIPS BETWEEN TEACHER EFFECT SCORES AND FINAL PAS PORTFOLIO RATINGS.

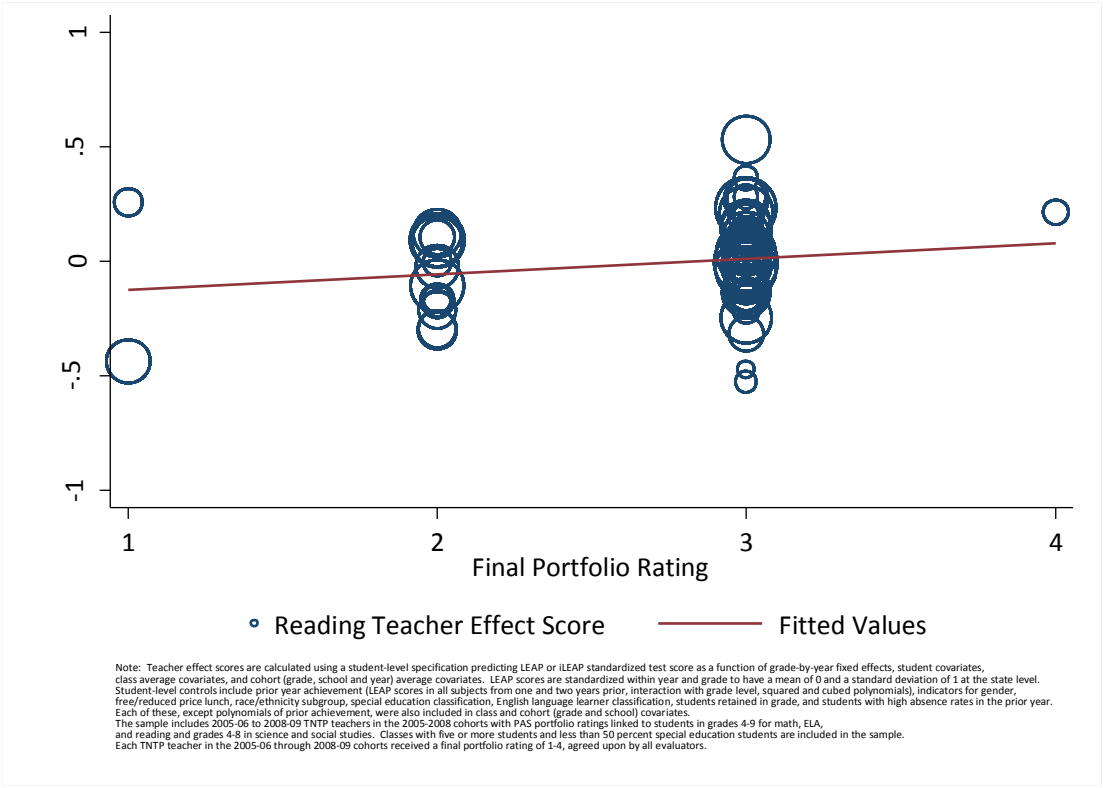


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

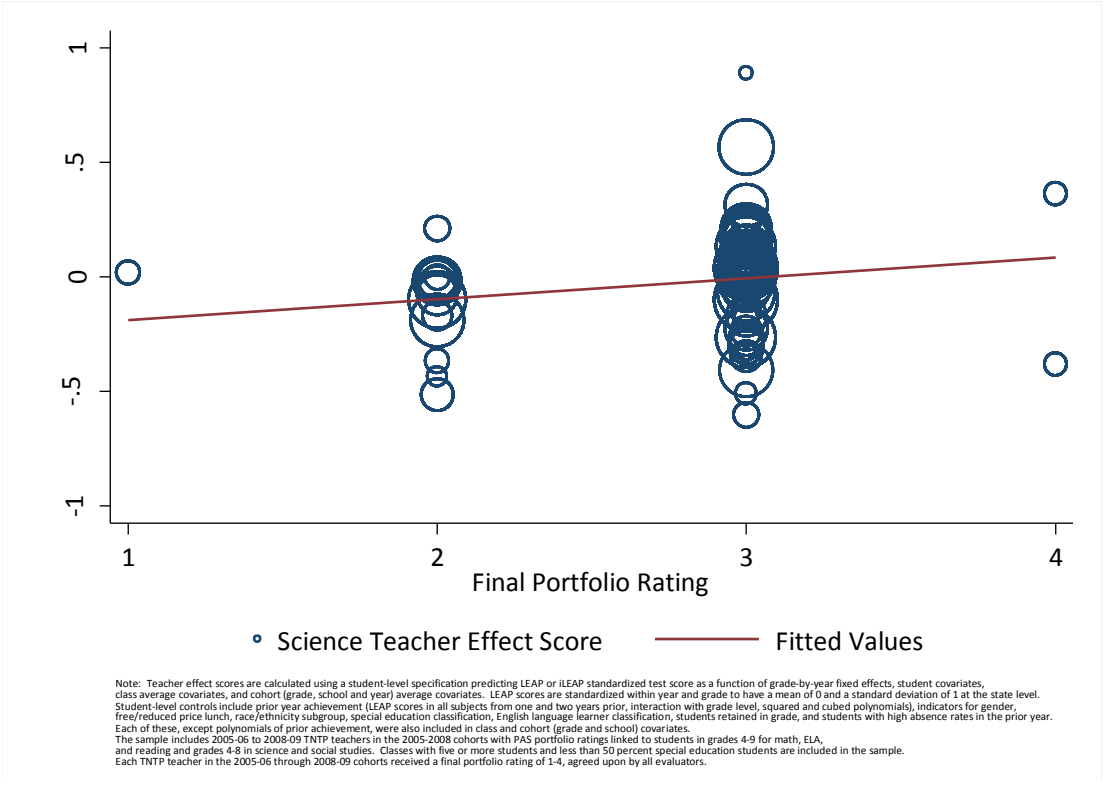


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates. The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample. Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

FIGURE 1B. RELATIONSHIPS BETWEEN TEACHER EFFECT SCORES AND FINAL PAS PORTFOLIO RATINGS.

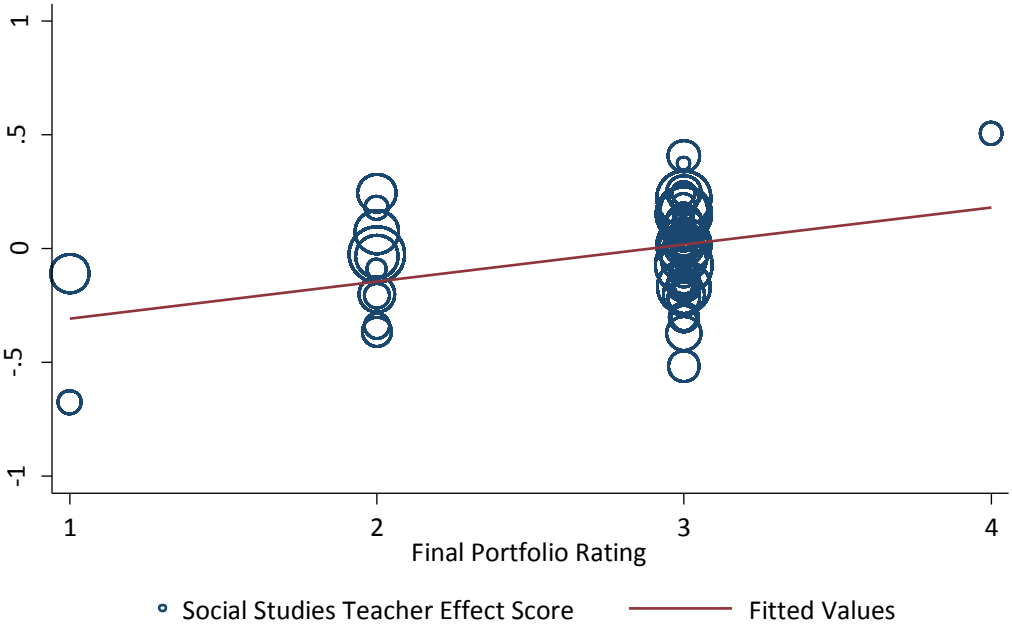


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

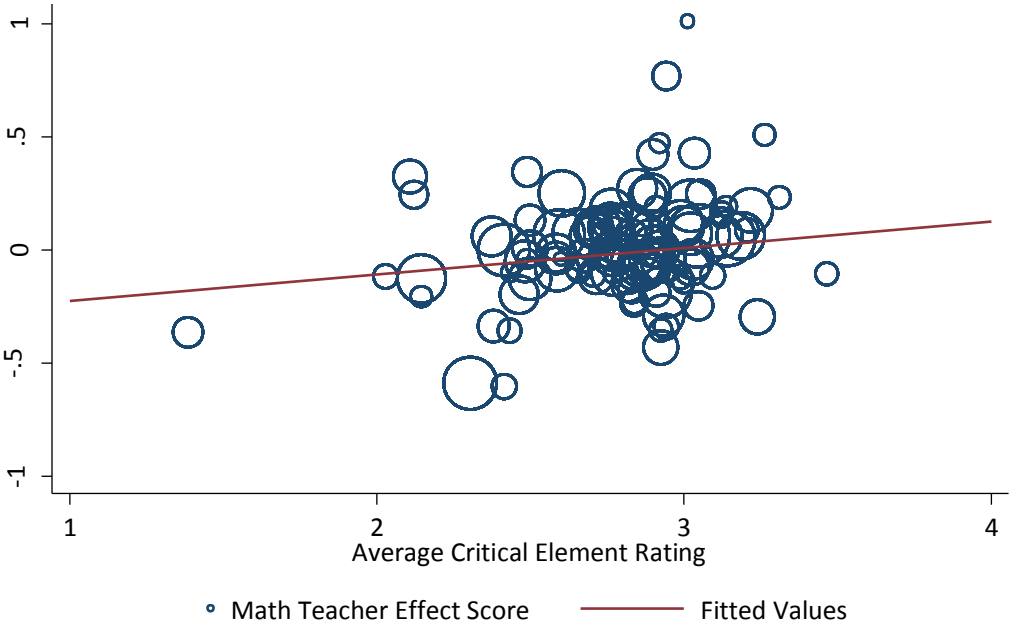


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

FIGURE 1C. RELATIONSHIPS BETWEEN TEACHER EFFECT SCORES AND FINAL PAS PORTFOLIO RATINGS.

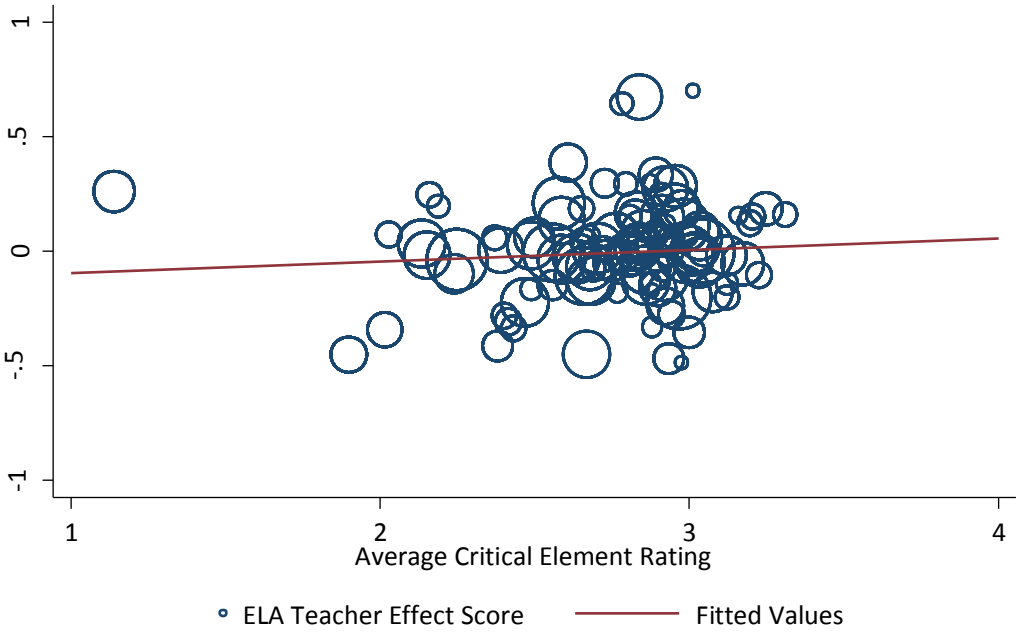


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
Each TNTP teacher in the 2005-06 through 2008-09 cohorts received a final portfolio rating of 1-4, agreed upon by all evaluators.

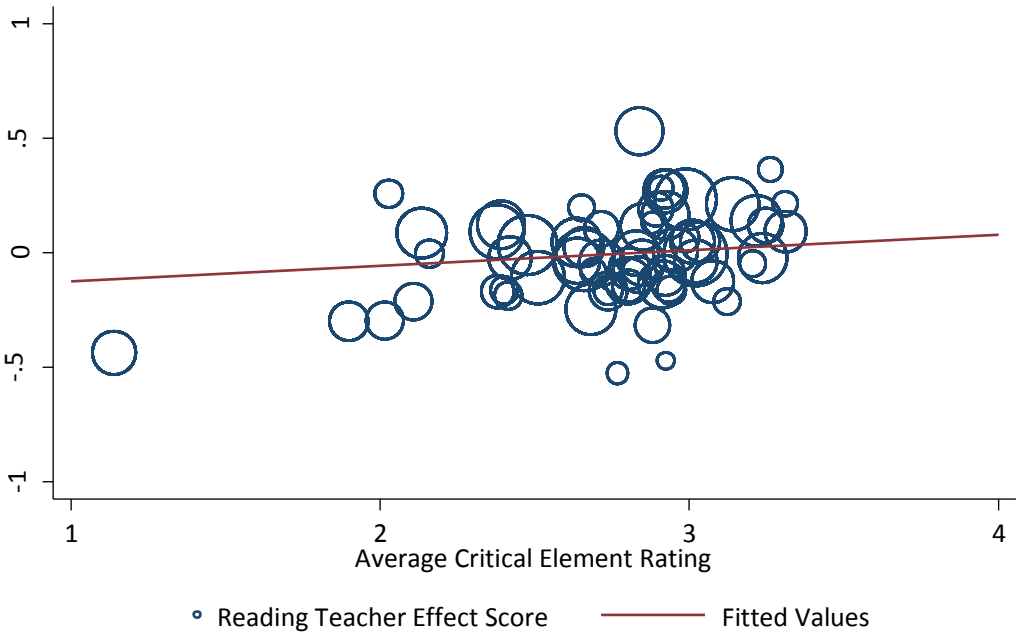


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

FIGURE 1D. RELATIONSHIPS BETWEEN TEACHER EFFECT SCORES AND FINAL PAS PORTFOLIO RATINGS

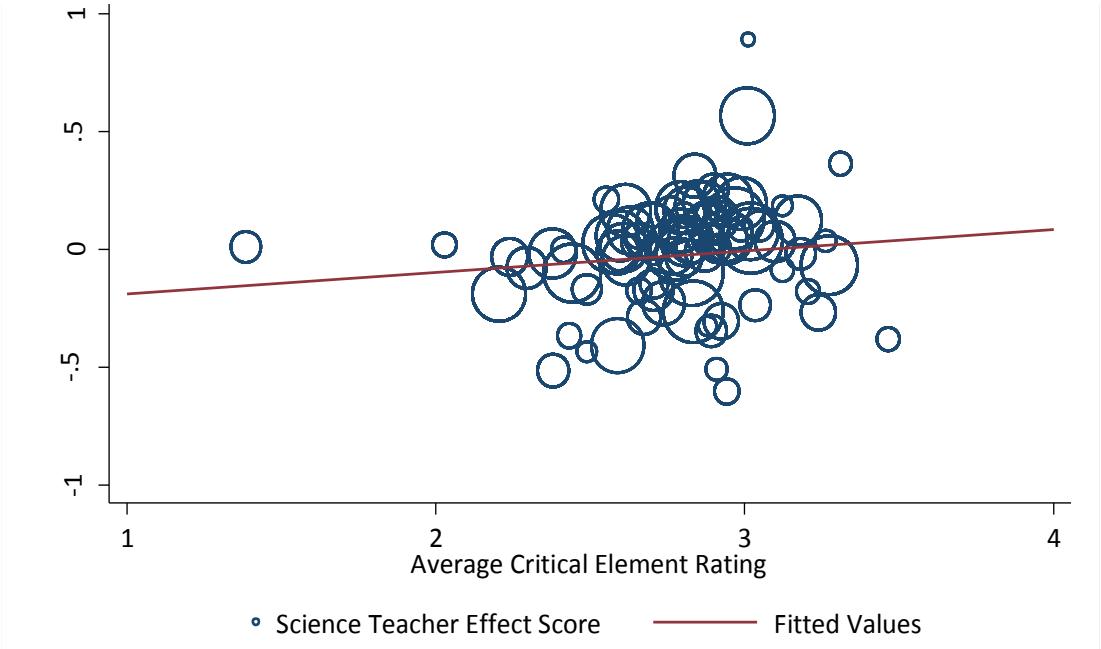


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

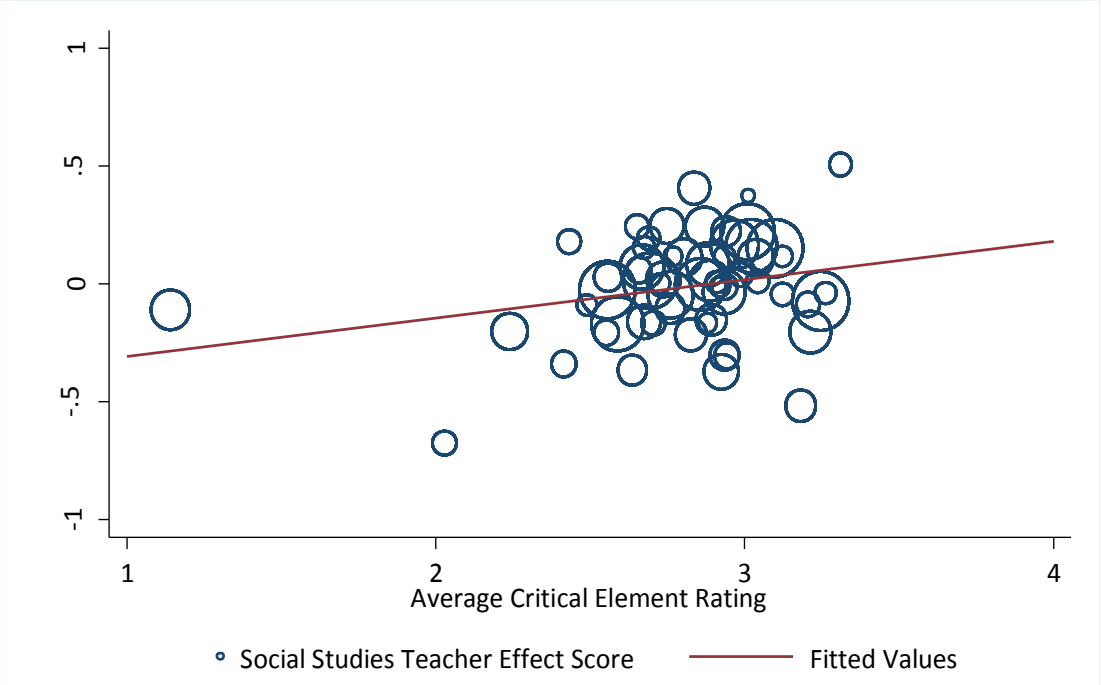


Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

FIGURE 1E. RELATIONSHIPS BETWEEN TEACHER EFFECT SCORES AND FINAL PAS PORTFOLIO RATINGS.



Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.



Note: Teacher effect scores are calculated using a student-level specification predicting LEAP or iLEAP standardized test score as a function of grade-by-year fixed effects, student covariates, class average covariates, and cohort (grade, school and year) average covariates. LEAP scores are standardized within year and grade to have a mean of 0 and a standard deviation of 1 at the state level. Student-level controls include prior year achievement (LEAP scores in all subjects from one and two years prior, interaction with grade level, squared and cubed polynomials), indicators for gender, free/reduced price lunch, race/ethnicity subgroup, special education classification, English language learner classification, students retained in grade, and students with high absence rates in the prior year. Each of these, except polynomials of prior achievement, were also included in class and cohort (grade and school) covariates.
The sample includes 2005-06 to 2008-09 TNTP teachers in the 2005-2008 cohorts with PAS portfolio ratings linked to students in grades 4-9 for math, ELA, and reading and grades 4-8 in science and social studies. Classes with five or more students and less than 50 percent special education students are included in the sample.
The average critical element score is the mean of all critical element ratings across all sources and evaluators for each teacher with a final portfolio rating of 1-4 between 2005-06 and 2008-09.

Appendices

Parent/Guardian Survey

Teacher_____ **School**_____

NOTE TO PARENTS/GUARDIANS:
Please fill out the following questions about your child's teacher. When finished, please sign the survey and return it to the teacher in a sealed envelope with your signature across the seal of the envelope flap. Thank you for your time.

For questions 1-6 please fill in the circle that most reflects your opinions:

	Always	*Usually*	*Sometimes*	*Never*	*I Don't Know*
1. Does your child feel good about being in this teacher's class?	❍	❍	❍	❍	❍
2. Does your child feel challenged by the work in this class?	❍	❍	❍	❍	❍
3. Does your child seem interested in what he/she is learning?	❍	❍	❍	❍	❍
4. Do you know what your child is learning in this class?	❍	❍	❍	❍	❍
5 Do you know how your child is performing in this class?	❍	❍	❍	❍	❍
6. Does this teacher behave professionally toward you and your child?	❍	❍	❍	❍	❍

For questions 7-12, please answer each question and explain your answers to the best of your abilities.

7. Has this teacher contacted you about your child? ___Yes ___No

8. Would you feel comfortable calling or talking to this teacher about your child's work?

___Yes ___No

9. Please explain:

10. Please rate your overall satisfaction with this teacher:

___Very Satisfied __ Satisfied __ Unsatisfied __ Very Unsatisfied

11. Please explain:

12. This survey allowed me to express my feelings about this teacher in a helpful way.
__ Agree __Somewhat Agree __Somewhat Disagree __Disagree

Student Survey (Grades 4-12)

Teacher_____

Grade Level_____ Subject(s) _____

Survey Administrator_____

Teacher Directions:
Complete the requested information above and give this form and sufficient copies of the student survey to the survey administrator who will supervise your students as they complete surveys independently.

Survey Administrator Directions:
Thank you for assisting this teacher in surveying his/her students. These surveys are an important part of this teacher's PAS submission, which will be reviewed to determine the quality of his/her performance. Before administering this survey, please read the following five guidelines.

1. Before distributing the survey, read aloud the "Note to Students" below.

2. For some students — depending on age and ability levels — you may want to read aloud each question and its answer choices, and give students a specified response time. If you choose to administer the survey in this manner, it is recommended that you:

   *require students to put down their pencils/pens between questions; and
   *remind students to listen to all answer choices before making their choice. It may be necessary to repeat each question and its applicable answer choices for clarity and comprehension.

3. Students should complete the survey in silence (no talking or sharing).

4. The survey should take the students approximately 20 minutes to complete. However, use your judgment to determine if additional time is needed for all students to finish.

5. Place the completed surveys and this cover sheet in a sealed envelope; sign your name across the seal of the envelope and return it to the teacher.

Note to Students
"Your teacher, _____, is being evaluated. One of the best ways to evaluate him/her is to ask you how you feel he/she is doing as your teacher, what you have learned in this class, and how you think he/she can improve. For the next 20 minutes, answer the questions on this survey [show paper]. It is very important that you answer the questions honestly. Do not write your name on the survey. It is also important that you complete this survey quickly and quietly. If you finish early, please show respect to your classmates by staying quiet until they too have finished."

I hereby certify that I read the guidelines above and administered the survey accordingly.

Survey Administrator Signature _____ **Date** _____

Student Survey (Grades 4-12)

Name of Teacher_____

Directions: Answer questions 1-10 by filling in the circle that best answers each question. For questions 11 and 12, put a check in the blank that best answers each question.

	Always	*Usually*	*Sometimes*	*Not Very Often*	*Never*
1. Does this teacher challenge you and expect a lot from you?	❍	❍	❍	❍	❍
2. Is this teacher fair with the classroom rules and procedures?	❍	❍	❍	❍	❍
3. Is this teacher's classroom a good environment for learning?	❍	❍	❍	❍	❍
4. Do you have to work hard to succeed in this class?	❍	❍	❍	❍	❍
5. Does this teacher help you when you need it or when you don't understand something?	❍	❍	❍	❍	❍
6. Do you think what you learn in this class is important?	❍	❍	❍	❍	❍
7. Does this teacher know when you are working hard and doing your best work?	❍	❍	❍	❍	❍
8. Is this teacher fair when grading you?	❍	❍	❍	❍	❍
9. Do you know how you are performing (your grade) in this class?	❍	❍	❍	❍	❍
10. Does this teacher help students outside of class?	❍	❍	❍	❍	❍

11. I think that he/she is a good teacher.

__ Agree __Somewhat Agree __Somewhat Disagree __Disagree

Is there anything else you would like us to know about this teacher?

12. This survey allowed me to express my feelings about this teacher in a helpful way.

__ Agree __Somewhat Agree _Somewhat Disagree __Disagree

Practitioner Teacher's name: _____

Practitioner Teacher's school: _____

Louisiana Practitioner Teacher Program

End-of-Year Principal Survey

We appreciate you taking the time to fill out this survey. The information you provide on this survey is very important as it will be used in the evaluation process of this beginning teacher. Your responses will offer insight about the effectiveness of this new teacher. In addition, your responses will allow us to learn about your experiences with teachers enrolled in our Practitioner Teacher Program (PTP) and help us to improve the program. Please remember, you may choose to skip any questions you don't wish to answer. After completing this survey please place it in an envelope, seal it, and sign your name across the back flap of the envelope. **Return the sealed, signed envelope to the Practitioner Teacher.** If you have any questions about this survey, feel free to call Nicole Bono, Instruction/Assessment Support Manager, LPTP at 225-644-2240.

Please check [✓] your responses to the survey questions below.

1. How you would compare this first year PTP teacher to all first year teachers you have worked with this year in terms of raising student achievement?

❍ Much better than other first year teachers
❍ Better than other first year teachers
❍ About the same as other first year teachers
❍ Worse than other first year teachers
❍ Much worse than other first year teachers

If Middle or High School:

2. Did this teacher raise student achievement by at least one grade level?

❍ Yes ❍ No

If Elementary:

	Yes	No	N/A
3. Did this teacher raise student achievement by at least one grade level in ELA (English Language Arts/Literacy)?	❍	❍	❍
4. Did this teacher raise student achievement by at least one grade level in MATH?	❍	❍	❍

5. This PTP Teacher defined, created, and enforced high standards for student behavior.

❍ Strongly Agree ❍ Agree ❍ Somewhat Agree
❍ Somewhat Disagree ❍ Disagree ❍ Strongly Disagree ❍ Not Sure

6. This PTP Teacher has designed and delivered standards-based instruction appropriate for the general student population.

❍ Strongly Agree ❍ Agree ❍ Somewhat Agree
❍ Somewhat Disagree ❍ Disagree ❍ Strongly Disagree ❍ Not Sure

7. This PTP Teacher has used assessment results to guide instructional choices.

❍ Strongly Agree ❍ Agree ❍ Somewhat Agree
❍ Somewhat Disagree ❍ Disagree ❍ Strongly Disagree ❍ Not Sure

8. This PTP Teacher has established positive, professional relationships with students, parents/guardians, and school colleagues.

❍ Strongly Agree ❍ Agree ❍ Somewhat Agree
❍ Somewhat Disagree ❍ Disagree ❍ Strongly Disagree ❍ Not Sure

9. Are there any additional comments you would like to make about this teacher?

10. Overall, do you believe that PTP Teachers (Teach For America/Teach Baton Rouge/teachNOLA) in your school have made a positive difference in the school environment?

❍ Yes ❍ No

11. Overall, how satisfied are you with the PTP Teachers (Teach For America/Teach Baton Rouge/teachNOLA) in your school?

❍ Very Satisfied ❍ Satisfied ❍ Somewhat Satisfied
❍ Somewhat Dissatisfied ❍ Dissatisfied ❍ Very Dissatisfied

12. Would you hire PTP Teachers again?

❍ Yes ❍ No

13. Is your school ☐ an elementary school ☐ a middle school ☐ a high school

PAS Rubrics: Classroom Environment/Culture

CRITICAL ELEMENT	VIDEOTAPE	SURVEYS	PD OBSERVATIONS
Establishes systems and routines to maximize instructional time	☐ Teacher and students perform class routines and transitions between activities with minimal impact on instructional time. Ex E P I		☐ Positive evidence that the teacher "implements and practices timesaving procedures." (Executes Effectively) Ex E P I
Defines, communicates, and enforces high standards for student behavior	☐ Teacher enforces behavioral standards consistently and appropriately. ☐ Students interact with each other and teacher in an appropriate manner. Ex E P I	☐ STUDENTS: "Always/ Usually"as a majority of responses to question 2 ☐ PRINCIPAL "Agree/ Somewhat Agree" as a response to question 5 Ex E P I	☐ Positive evidence that the teacher "communicates behavioral expectations by teaching and practicing rules & consequences and reacting swiftly, fairly, assertively & consistently." (Executes Effectively) Ex E P I
Creates a physical environment conducive to learning	☐ Classroom physical layout facilitates routines/ transitions and encourages student learning. Ex E P I		
Promotes a positive learning environment for all students	☐ Teacher ensures that students are engaged in lesson or learning activities. Ex E P I	☐ PARENTS: "Always/ Usually" as a majority of responses to question 1 ☐ STUDENTS: "Always/ Usually" as a majority of responses to question 3 Ex E P I	☐ Positive evidence that the teacher "creates a welcoming classroom ☐ environment to instill values (respect, tolerance, kindness, collaboration)." (Invests Students) Ex E P I

PAS Rubrics – Global Version - Grades 4-12

PAS Rubrics: Classroom Instructional Design and Delivery

CRITICAL ELEMENT	INSTRUCTIONAL UNIT	VIDEOTAPE	SURVEYS	PD OBSERVATIONS
Plans in a backwards design way to set standards-based instructional goals that reflect high expectations for all students	☐ Context report shows unit link to Louisiana State or district-specific standards. ☐ Priority levels assigned to GLE's in the context report are logical/reasonable. Ex E P I		☐ PARENTS: "Always/ Usually" as a majority of responses to question 2 ☐ STUDENTS: "Always/ Usually" as a majority of responses to question 4 Ex E P I	☐ Positive evidence that the teacher develops "standards-aligned, measurable, ambitious, & feasible goals." (Sets Big Goals) ☐ Positive evidence that the teacher "Backwards-plans by breaking down longerterm goals into bundles of objectives and mapping them across the school year. " (Plans Purposefully) Ex E P I
Designs and delivers instruction appropriate for general student population	☐ Instructional strategies in lesson plans are appropriate for the content matter/grade level and general student population. ☐ Lesson plan activities align with standards/ GLE's, and will Ex E P I	☐ Majority of students are responsive to teacher's instructional strategies. Ex E P I	☐ PRINCIPAL: "Agree/Somewhat Agree" as a response to question 6 Ex E P I	☐ Positive evidence that the teacher "Creates rigorous, objective-driven lesson plans." (Plans Purposefully) Ex E P I
Modifies instruction to meet the individual needs of diverse learners	☐ Instructional strategies in lesson plans are varied and include modifications for students with individual learning differences or needs. Ex E P I	☐ Individual students are engaged and able to participate in lesson and learning activities. Ex E P I	☐ STUDENTS: "Always/ Usually" as a majority of responses to question 5 Ex E P I	☐ Positive evidence that the teacher "differentiates instructional plans for individual students." (Plans Purposefully) Ex E P I
Orders and structures lessons that promote student retention of content knowledge	☐ Unit Plan Outline objectives are logically sequenced so that concepts/ skills build on and reinforce one another. ☐ Unit Plan Outline objectives align with standards/GLE's that are listed Ex E P I	☐ Teacher sets up and creates logical connections between lesson plan parts/segments. Ex E P I		
Delivers instruction that encourages student engagement	☐ Activities and materials used initiate and maintain student interest in lesson or learning activities. Ex E P I	☐ Students are engaged and invested in com-pleting lesson or learning activities Ex E P I	☐ STUDENTS: "Always/ Usually" as a majority of responses to question 6 Ex E P I	

PAS Rubrics – Global Version - Grades 4-12

PAS Rubrics: Assessment

CRITICAL ELEMENT	INSTRUCTIONAL UNIT	DSA REPORT	SURVEYS	PD OBSERVATIONS
Selects and administers assessments appropriately	☐ Assessments used in lesson plans and student work samples are appropriate for the general student population. ☐ Assessments in student work samples align with the P1 or P2 standard that PT has selected to track Ex E P I	☐ Assessments described or documented in report are appropriate for curricular goals and general student population. Ex E P I	☐ STUDENTS: "Always/ Usually" as a majority of responses to question 7 Ex E P I	☐ Positive evidence that the teacher "Creates or obtains diagnostics and daily, periodic and annual assessments." (Plans Purposefully) Ex E P I
Monitors student progress with a variety of assessments that address differing learning styles	☐ Assessments in lesson plans and student work samples allow individual students to demonstrate learning in a variety of ways. Ex E P I	☐ Assessments described or documented in report are varied. Ex E P I		
Modifies instructional strategies based on assessment results	☐ All progress reports describe modifications in instruction in order to meet individual student needs and ensure progress toward curricular goals. Ex E P I	☐ Report describes ways teacher has modified instructional strategies based on assessment findings. Ex E P I	☐ PRINCIPAL: "Agree/ Somewhat Agree" as a response to question 7 Ex E P I	
Communicates student performance levels and strategies for progress to student and parents/ guardians	☐ Student work samples communicate performance levels and strategies for improvement to students. Ex E P I		☐ PARENTS: "Always/ Usually" as majority of responses to q 5 ☐ STUDENTS: "Always/ Usually" as majority of responses to q 9 Ex E P I	☐ Positive evidence that the teacher "Checks for understanding frequently by questioning and listening, and provides feedback (that affirms right answers and corrects wrong answers)." (Executes Effectively) Ex E P I

PAS Rubrics – Global Version - Grades 4-12

PAS Rubrics: Professionalism

CRITICAL ELEMENT	VIDEOTAPE	SURVEYS	PD OBSERVATIONS	CSL FORM
Reflects on and revises practice continuously to improve teaching performance	☐ Lesson Analysis reflects critically on teaching strengths and areas for improvement as shown in the videotaped segment. Ex E P I		☐ Action Report reflects critically on PD feedback and describes specific actions or steps taken to improve classroom performance. Ex E P I	
Establishes positive, professional relationships with student, parents, and colleagues	☐ Teacher maintains positive rapport with students without compromising role as instructional leader. Ex E P I	☐ PARENTS: "Always" /"Usually" as a majority of responses to question 6 ☐ PARENTS: Most parents answer "yes" to question 8 ☐ PRINCIPAL: "Agree/Somewhat Agree" as a response to question 8 Ex E P I		
Fulfills professional responsibilities				☐ Teacher successfully meets seminar requirements and expectations (the "pass"box is checked). Ex E P I

PAS Rubrics – Global Version - Grades 4-12

PAS Rubrics: Student Achievement

CRITICAL ELEMENT	INSTRUCTIONAL UNIT	VIDEOTAPE	DSA REPORT	SURVEYS
Student behavior and performance indicate progress toward or achievement of curricular goals	☐ Student work samples and progress reports indicate that the majority of 6 students achieved progress towards mastering selected P1 or P2 standard. Ex E P I	☐ Most students demonstrate comprehension of lesson objective and content/skill being taught. Ex E P I	☐ Report indicates that the teacher has increased student academic achievement. Ex E P I	☐ STUDENTS: "Always/Usually" as a majority of responsesto question 1 ☐ PRINCIPAL: Rates teacher "Better or Much Better" than other first year teachers on question 1 & a yes response to question 2, 3, or 4. Ex E P I

PAS Rubrics – Global Version - Grades 4-12

PAS Synthesis Form

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Establishes systems and routines to maximize instructional time	Videotape		M	Ex	CLASSROOM ENVIRONMENT/ CULTURE
	PD Observations		L		
Defines, communicates, and enforces high standards for student behavior	Videotape		M	E	
	Surveys		M		
	PD Observations		L	P	
Creates a physical environment conducive to learning	Videotape		M	I	
Promotes a positive learning environment for all students	Videotape		M		
	Surveys		M		
	PD Observations		L		

PAS Synthesis Form

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Plans in a backwards- design way to set standards-based instructional goals that reflect high expectations for all students	Instructional Unit		H	Ex	INSTRUCTIONAL DESIGN AND DELIVERY
	Surveys		M		
	PD Observations		L		
Designs and delivers instruction appropriate for general student population	Instructional Unit		H	E	
	Videotape		M		
	Surveys		M		
	PD Observations		L		
Modifies instruction to meet the individual needs of diverse learners	Instructional Unit		H	P	
	Videotape		M		
	Surveys		M		
	PD Observations		L		
Orders and structures lessons that promote student retention of content knowledge	Instructional Unit		H	I	
	Videotape		M		
Delivers instruction that encourages student engagement	Instructional Unit		H		
	Videotape		M		
	Surveys		M		

PAS Synthesis Form

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Selects and administer assessments appropriately	Instructional Unit		H	Ex	ASSESSMENT
	DSA Report		L		
	Surveys (grades 4-12 only)		M		
	PD Observations		L		
Monitors student progress with a variety of assessments that address differing learning styles	Instructional Unit		H	E	
	DSA Report		L		
Modifies instructional strategies based on assessment results	Instructional Unit		H	P	
	DSA Report		L		
	Surveys		M		
Communicates student performance levels and strategies for progress toward curricular goals to students and parents/guardians	Instructional Unit		H	I	
	Surveys		M		
	PD Observations		L		

PAS Synthesis Form

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Reflects on and revises practice continuously to improve teaching performance	Videotape		M	Ex E P I	PROFESSIONALISM
	PD Observations		L		
Establishes positive, professional relationships with students, parents/ guardians, and colleagues	Videotape		M		
	Surveys		M		
Fulfills professional responsibilities	CSL Form		L		

CRITICAL ELEMENT	COMPONENT	RATING	SIG.	FRAMEWORK AREA RATING	
Student behavior and performance indicate progress toward or achievement of curricular goals	Instructional Unit		H	Ex E P I	STUDENT ACHIEVEMENT
	Videotape		M		
	DSA Report		L		
	Surveys		M		

APPENDIX D. DESCRIPTION OF COHORTS NOT INCLUDED IN PRIMARY DATA ANALYSIS

While we have data available for both the 2004-05 and 2009-10 cohorts, they are excluded from the analysis presented above. Each of these two cohorts is excluded for a different reason, with the reasons described in this appendix.

Hurricane Katrina has a massive impact on the state of Louisiana, especially New Orleans and its surroundings. As a result of this natural disaster many students left the New Orleans public schools by choice or through relocation programs in other states. The result of this disruption is that many student test scores are missing for this cohort. Further complicating any potential analysis involving this cohort is that there are fewer teachers, only 79, in this cohort.

As noted in footnote 16, data from the 2009-10 cohort is excluded from this report's analysis because the final portfolio scoring for this cohort is different than for the earlier cohorts. Teachers in the four earlier cohorts received integer scores of 1, 2, 3, or 4 for their final portfolio score whereas teachers in the 2009-10 cohort were assigned a dichotomous pass or fail portfolio score.

Teach For America / Teach Baton Rouge / teachNOLA
2008-09 PAS Observation Feedback Form

Practitioner Teacher: _____ *Date: _____*
Program Director/Observation Specialist: _____

PAS Evidence:

SETS BIG GOALS
Develop a standards-aligned, measurable, ambitious and feasible goal.
INVESTS STUDENTS
Create a welcoming classroom environment to instill values (respect, tolerance, kindness, collaboration).
PLANS PURPOSEFULLY
Create or obtain diagnostics and daily, periodic and annual assessments
Backwards-plan by breaking down longer-term goals into bundles of objectives and mapping them across the school year
Create rigorous, objective-driven lesson plans
Differentiate instructional plans for individual students based on their unique learning profiles.
EXECUTES EFFECTIVELY
Check for understanding frequently by questioning and listening, and provide feedback (that affirms right answers and corrects wrong answers).
Communicate behavioral expectations by teaching and practicing rules and consequences, and reacting swiftly, fairly, assertively and consistently.
Implement and practice time-saving procedures (for transitions, dissemination and collection of supplies or homework, etc.).

2009-10 PAS Observation Feedback Form
Teach For America / Teach Baton Rouge / teachNOLA/LTF

Practitioner Teacher: _____ Date: _____

Program Director/Observation Specialist: _____

1) *Pre-Novice – does not demonstrate the action and/or cannot articulate key ideas/strategies/criteria*
2) *Novice – is able to describe the action and its importance to student learning, and attempts the action.*
3) *Beginning Proficiency – makes a solid impact by technically executing the basic action.*
4) *Advanced Proficiency – makes a significant impact by implementing strategies with their purpose in mind.*
5) *Exemplary – makes a dramatic impact by thinking beyond the boundaries of traditional methods, stopping at nothing to yield the sought-after results.*

Teacher Action	Level of Proficiency
SETS BIG GOALS	
Develop a standards-aligned, measurable, ambitious and feasible goal **	
INVESTS STUDENTS	
Develop students' **rational understanding that they can achieve by working hard ("I can")**	
Develop students' **rational understanding that they will benefit from achievement ("I want")**	
Employ **appropriate role models**	
Consistently **reinforce students' efforts toward the big goal**	
Create a **welcoming classroom environment** to instill values (respect, tolerance, kindness, collaboration) **	
Respectfully mobilize students' influencers (e.g., family, peers, coach, pastor, etc)	
PLANS PURPOSEFULLY	
Create or obtain diagnostics and daily, periodic and annual assessments	
Backwards-plan by breaking down longer-term goals into bundles of objectives and mapping them across the school year	
Create rigorous, **objective-driven lesson plans**	
Differentiate instructional plans for individual students based on their unique learning profiles **	
Establish age-appropriate long- and short-term behavioral management plans	
Design classroom procedures (for transitions, collecting and handing out papers, taking roll, grading, etc.)	
EXECUTES EFFECTIVELY	
Clearly **present academic content** so that students comprehend key information and ideas	
Facilitate, manage, and **coordinate student practice** (in differentiated ways, if necessary)	
Check for understanding frequently by questioning and listening, **and provide feedback** (that affirms right answers and corrects wrong answers) **	
Communicate behavioral expectations by teaching and practicing rules and consequences, and reacting swiftly, fairly, assertively and consistently **	
Implement and practice **time-saving procedures** (for transitions, dissemination and collection of supplies or homework, etc.) **	
Evaluate and keep track of students' performance on assessments	
CONTINUOUSLY INCREASES EFFECTIVENESS	
Gauge progress and notable gap(s) between student achievement and big goals by examining assessment data	
Identify the student habits most influencing progress and gaps between student achievement and big goals	
Isolate the teacher actions contributing to key aspects of students performance by gathering data and reflecting on teacher performance	
Identify the underlying factors (e.g., knowledge, skill, mindset, dilemma, bias, emotion, etc.) causing teacher action	
Access relevant learning experiences that direct and inform teacher improvement	
After a cycle of data collection, reflection and learning, **adjust course** as necessary	
WORKS RELENTLESSLY	
Persist in the face of considerable challenges	
Pursue and secure additional instructional time and resources	
Sustain the intense energy necessary to reach the ambitious big goals	

**** Denotes action or behavior PD should comment on below for PAS purposes.**

PAS Evidence:

SETS BIG GOALS
Develop a standards-aligned, measurable, ambitious and feasible goal
INVESTS STUDENTS
Create a welcoming classroom environment to instill values (respect, tolerance, kindness, collaboration).
PLANS PURPOSEFULLY
Create or obtain diagnostics and daily, periodic and annual assessments **Backwards-plan by breaking down longer-term goals into bundles of objectives and mapping them across the school year** **Create rigorous, objective-driven lesson plans** **Differentiate instructional plans for individual students based on their unique learning profiles.**
EXECUTES EFFECTIVELY
Check for understanding frequently by questioning and listening, and provide feedback (that affirms right answers and corrects wrong answers). **Communicate behavioral expectations by teaching and practicing rules and consequences, and reacting swiftly, fairly, assertively and consistently.** **Implement and practice time-saving procedures (for transitions, dissemination and collection of supplies or homework, etc.).**

APPENDIX I: SUMMARY STATISTICS OF THE DISTRIBUTION OF FINAL PORTFOLIO SCORES, BY SUBJECT AREA (2005-06 THROUGH 2008-09)

Math

FINAL PORTFOLIO RATING	FREQUENCY	PERCENT	CUMULATIVE
1	1	0.92	0.92
2	14	12.84	13.76
3	91	83.49	97.25
4	3	2.75	100.00
TOTAL	109	100.00	

ELA

FINAL PORTFOLIO RATING	FREQUENCY	PERCENT	CUMULATIVE
1	2	1.87	1.87
2	22	20.56	22.43
3	82	76.64	99.07
4	1	0.93	100.00
TOTAL	107	100.00	

Reading

FINAL PORTFOLIO RATING	FREQUENCY	PERCENT	CUMULATIVE
1	2	3.13	3.13
2	13	20.31	23.44
3	48	75.00	98.44
4	1	1.56	100.00
TOTAL	64	100.00	

Science

FINAL PORTFOLIO RATING	FREQUENCY	PERCENT	CUMULATIVE
1	1	1.20	1.20
2	13	15.66	16.87
3	67	80.72	97.59
4	2	2.41	100.00
TOTAL	83	100.00	

Social Studies

FINAL PORTFOLIO RATING	FREQUENCY	PERCENT	CUMULATIVE
1	2	3.39	3.39
2	10	16.95	20.34
3	46	77.97	98.31
4	1	1.69	100.00
TOTAL	59	100.00	

APPENDIX J: SUMMARY STATISTICS OF ALL FRAMEWORK SCORES, AS MEASURED BY THE AVERAGE OF ALL CRITICAL ELEMENTS, ACROSS ALL SOURCES AND ASSESSORS, BY SUBJECT AREA (2005-06 THROUGH 2008-09)

Math

Average of All Critical Element Scores, Across Scores and Assessors

	PERCENTILES	SMALLEST		
1%	2.119565	1.386364		
5%	2.30303	1.386364		
10%	2.462121	1.386364	OBS	4735
25%	2.673913	1.386364	SUM OF WGT.	4735
50%	2.829545		MEAN	2.796818
		LARGEST	STD. DEV.	.2583789
75%	2.940476	3.463768		
90%	3.119565	3.463768	VARIANCE	.0667597
95%	3.184783	3.463768	SKEWNESS	-.9523445
99%	3.238636	3.463768	KURTOSIS	3 5.184248

ELA

Average of All Critical Element Scores, Across Scores and Assessors

	PERCENTILES	SMALLEST		
1%	1.897727	1.141304		
5%	2.152174	1.141304		
10%	2.25	1.141304	OBS	4040
25%	2.586957	1.141304	SUM OF WGT.	4040
50%	2.806818		MEAN	2.74002
		LARGEST	STD. DEV.	.3049365
75%	2.956522	3.311594		
90%	3.045455	3.311594	VARIANCE	.0929863
95%	3.121212	3.311594	SKEWNESS	-1.273762
99%	3.246377	3.311594	KURTOSIS	6.229585

Reading

Average of All Critical Element Scores, Across Scores and Assessors

	PERCENTILES	SMALLEST		
1%	1.141304	1.141304		
5%	2.108696	1.141304		
10%	2.378788	1.141304	OBS	2271
25%	2.636364	1.141304	SUM OF WGT.	2271
50%	2.829545		MEAN	2.75428
		LARGEST	STD. DEV.	.3617257
75%	2.988636	3.311594		
90%	3.141304	3.311594	VARIANCE	.1308455
95%	3.246377	3.311594	SKEWNESS	-1.363744
99%	3.311594	3.311594	KURTOSIS	6.454396

Science

Average of All Critical Element Scores, Across Scores and Assessors

	PERCENTILES	SMALLEST		
1%	2.204545	1.386364		
5%	2.295455	1.386364		
10%	2.442029	1.386364	OBS	4070
25%	2.630435	1.386364	SUM OF WGT.	4070
50%	2.811594		MEAN	2.783063
		LARGEST	STD. DEV.	.2576717
75%	2.940476	3.463768		
90%	3.072723	3.463768	VARIANCE	.0663947
95%	3.206522	3.463768	SKEWNESS	-.727768
99%	3.275362	3.463768	KURTOSIS	5.295745

Social Studies

Average of All Critical Element Scores, Across Scores and Assessors

	PERCENTILES	SMALLEST		
1%	2.028986	1.141304		
5%	2.416667	1.141304		
10%	2.555556	1.141304	OBS	2146
25%	2.688406	1.141304	SUM OF WGT.	2146
50%	2.869565		MEAN	2.833556
		LARGEST	STD. DEV.	.2888607
75%	3.01087	3.311594		
90%	3.206522	3.311594	VARIANCE	.0834405
95%	3.246377	3.311594	SKEWNESS	-1.929848
99%	3.261364	3.311594	KURTOSIS	12.07607

Center for Education Policy Research

HARVARD UNIVERSITY

50 Church Street, 4th Floor
Cambridge, MA 02138

www.gse.harvard.edu/cepr
cepr@gse.harvard.edu