Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Strunk, Katharine O.; Weinstein, Tracey L.; Makkonen, Reino
Sorting Out the Signal: Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals?
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 22, 2014, pp. 1-41
Arizona State University
Arizona, Estados Unidos

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 22 Number 100 November 10th, 2014 ISSN 1068-2341

Sorting Out the Signal: Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals?

Katharine O. Strunk
University of Southern California

Tracey L. Weinstein
StudentsFirst
&
Reino Makkonen
WestEd

Citation: Strunk, K., Weinstein, T., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives, 22*(100). http://dx.doi.org/10.14507/epaa.v22.1590

Abstract: There is increasing policy interest in the use of standards-based multiple measure teacher evaluation systems that include both observational and value-added measures of teacher effectiveness. The growing literature that assesses the relationships between these measures does so mainly in academic settings using a validity lens. While valuable in their own right, this evidence from research-based settings provides little evidence about how teachers and principals receive the different signals from multiple measures of effectiveness when implemented in district contexts. Using pairwise correlations and a series of ordinary least squares regressions, this study assesses the relationships between value-added measures of teacher effectiveness and

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/3/2014
Revisions received: 5/7/2014
Accepted: 5/8/2014

an observational measure of teacher practice as implemented in a district's pilot of a new standards-based multiple-measure teacher evaluation system. We find moderate correlations between value-added and observation-based measures, indicating that teachers will receive similar but not entirely consistent signals from each performance measure. We conclude by highlighting considerations for districts working to develop and implement standards-based multiple-measure teacher evaluation systems.
Keywords: teacher evaluation; teacher effectiveness; teacher quality; measuring teacher practice; value-added.

Entendiendo las señales: ¿Son relevantes los sistemas de múltiples medidas de efectividad docente para brindar información coherente a docentes y directores?

Resumen: Cada vez hay más interés en el uso de los sistemas de medidas múltiples de evaluación docente basados en estándares con medidas de observación y de valor añadido de la eficacia docente. La creciente literatura que evalúa las relaciones entre estas medidas lo hace principalmente en el ámbito académico utilizando la lente de la validez. Si bien valiosa, la evidencia basada en la investigación proporciona poca información sobre cómo los profesores y directores reciben las diferentes señales de los múltiples medidas de efectividad cuando se implementa en contextos distritales. Utilizando correlaciones por pares y una serie de regresiones de mínimos cuadrados ordinarios este estudio evalúa la relación entre las medidas de valor agregado de la eficacia docente y una medida observacional de la práctica docente como se aplica en una prueba piloto en un distrito con un sistema de múltiples medidas. Encontramos correlaciones moderadas entre el valor agregado y las medidas basadas en la observación, lo que indica que los docentes recibían señales enteramente coherentes similares pero no de cada medida de desempeño. Concluimos resaltando consideraciones para los distritos que trabajan para desarrollar e implementar sistemas de evaluación docente de medidas múltiples basadas en estándares.
Palabras clave: evaluación docente; efectividad docentes; calidad de los maestros; medición de la práctica docente; valor añadido.

Compreendendo os sinais: São relevantes os modelos de medidas múltiplas da eficácia do professor para fornecer informações consistentes para professores e diretores?

Resumo: Há um crescente interesse em utilizar sistemas de medidas múltiplas de avaliação de professores baseada em padrões, que incluem tanto medidas observacionais e de valor agregado sobre a eficácia do professor. A crescente literatura acadêmica avaliando a relação entre essas medidas usa principalmente o lente de validade. Embora valiosa, as pesquisas baseadas em evidências fornece pouca informação sobre como professores e diretores recebem os diferentes sinais de medidas múltiplas de eficácia quando implementado em contextos distritais. Usando correlações em pares e uma série de regressões este estudo avalia a relação entre as medidas de valor agregado de eficácia do professor e uma medida de observação da prática docente aplicada em um teste piloto em um distrito escolar com um sistema de medidas múltiplas. Foram encontradas correlações moderadas entre as medidas de valor acrescentado baseados na observação, indicando que os professores receberam sinais inteiramente coerentes semelhantes, embora não de cada medida de desempenho. Concluímos, destacando considerações para os distritos que trabalham para desenvolver e implementar sistemas de avaliação de professores com base em padrões de medidas múltiplas.
Palavras-chave: avaliação docente; professores efetivos; qualidade dos professores; medição da prática docente; valor agregado.

Introduction

The problems with many current systems of teacher evaluation are now well-known. Not only are they often compliance-oriented, perfunctory and without consequence, but they produce little information that teachers and schools can use to help teachers improve or that schools and districts can draw upon when making personnel decisions such as promotion, retention and removal (Brandt, Mathers, Oliva, Brown-Simms, & Hess, 2007; Donaldson, 2009; Kauchak, Peterson, & Driscoll, 1985; Weisberg, Sexton, Mulhern, & Keeling, 2009). Given the problems with existing teacher evaluation systems and a growing understanding of the importance of high-quality teachers (see, for examples, Chetty, Friedman, & Rockoff, 2011; Goldhaber, Brewer, & Anderson, 1999; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), there has been increasing policy interest at the federal, state and local levels in the use of expanded teacher evaluation systems to assess and reward teacher effectiveness and to support the development of teachers' practice. To that end, nearly two-thirds of U.S. states have made changes to their teacher evaluation policies since 2009 in ways that require or encourage the use of revised, standards-based multiple-measure teacher evaluation systems (MMTES) (Jerald, 2012).

Such systems commonly require multiple measures of performance, including classroom observations, measures of teachers' contributions to their students' performance on standardized tests, and stakeholder surveys that measure parent and/or student beliefs about teacher quality. Yet to date, much of the evidence pertaining to the relationships among these measures of teacher effectiveness comes from studies of measures generated for academic research purposes and not from systems that are being implemented in districts for current or eventual stakes. As a result, despite federal and state incentives and policies that encourage or require the implementation of these new systems, there is as yet little empirical evidence regarding the usefulness of different measures to assess teachers' abilities to deliver high-quality instruction and improve student achievement in practice. Moreover, much of the extant research explores relationships between the two measures from a validity standpoint, attempting to assess if one measure is valid due to its correlation with other measures of effectiveness (e.g., Bell et al., 2012; Hill, Kapitula, & Umland, 2011; Holtzapple, 2003; Milanowski, 2004a, 2004b). There has been little research that specifically tackles the problem from the perspective of teachers and administrators implementing new evaluation systems, examining the different measures to see if the measures as they are given to teachers and principals in schools provide them with consistent signals about teacher effectiveness.

The lack of practice-based empirical evidence is particularly problematic for three reasons. First, the lessons that stem from studies in research settings cannot entirely speak to what will be found in contexts in which districts are actually implementing new systems for current or eventual stakes. In fact, we might imagine that we will see quite different relationships between observational and test-score based measures of effectiveness when districts implement their own systems. For example, students may not be assigned randomly to their classrooms (Rothstein, 2010), observers may be less well calibrated and trained in true-life situations than in research settings (TDOE, 2012), observers—and especially site administrators—may choose to or need to take other elements (such as good will between him/herself and the teacher, trust, etc.) into consideration when noting observations of "effectiveness" (Sartain, Stoelinga, & Brown, 2011), and political considerations may impact the content of observational frameworks or the methods used to generate test-score-based measures of effectiveness (frequently called value-added measures, or VAMs) (Anderson, 2012; Fleisher, 2012; TDOE, 2012). Given the myriad political, contextual and capacity realities facing districts that are implementing MMTES, the relationship between observation-based measures of effectiveness and VAMs found in research studies may not echo what will be found in districts.

Second, the observational protocols being discussed and implemented in MMTES tend to consist of multiple "standards" or "components," some of which can be observed during the actual classroom observation and some of which must be observed during pre- and post-conferencing and through other activities. Because much of the extant research—even those studies that stem from the implementation of actual MMTES—explores relationships between measures of effectiveness from a validity standpoint, most of the studies (e.g., Kane et al., 2011; Kane & Staiger, 2012; Mihaly et al., 2013) focus on the subset of information garnered from observation protocols that can theoretically be observed during a class session rather than across the whole observation rating. However, teachers and administrators receive the entire set of information as an overall score of practice-based effectiveness. As a result, the majority of previous studies do not address the consistency of the message about effectiveness between the measures as they are received in practice.

Third, applying a validity lens to the relationship between measures of effectiveness garnered from MMTES has caused some of the most influential research (e.g., Kane et al., 2013) to attempt to reduce the noise inherent in measures of teacher effectiveness to assess the relationship between the "true" or "underlying" measures. By adjusting measures such as VAMs to capture more "signal" and less "noise," policymakers can better understand whether or not non-test-score based measures of effectiveness are valid—that is, if they are associated with teachers' abilities to improve test scores. However, they do not provide much insight to districts and policymakers about the consistency of the signal that teachers and principals will receive about teacher effectiveness from systems that employ the various measures without statistical adjustments.

For these reasons, it is particularly important to study the relationships between the actual unadjusted measures used in teacher evaluation systems in practice. This study assesses the relationship between student test score-based measures of teacher effectiveness (VAMs) and an observation-based measure of teacher effectiveness, both of which are part of the Los Angeles Unified School District's (LAUSD) Initial Implementation Phase (IIP) of a new MMTES, called the Educator Growth and Development Cycle (EGDC). Using VAMs (called "Academic Growth over Time," AGT, in LAUSD) and classroom observation scores from the approximately 200 teachers who enrolled in the district's IIP and who had AGT data, we answer two questions about the relationships between these measures and between specific instructional practices captured in the classroom observation and AGT: (1) When implemented as a part of a standards-based multiple-measure teacher evaluation system, do value-added and observational measures of teacher effectiveness provide teachers and principals with a consistent signal of teacher effectiveness?; and (2) Do specific classroom practices measured by a district-generated observation rubric capture differences in value-added measures of effectiveness?

LAUSD provides a particularly interesting setting in which to ask these questions. First, it is the second largest school district in the country, serving nearly 700,000 students with approximately 25,000 teachers in 763 schools. Given the size of the district, its pilot was able to include over 400 teachers and over 100 school site administrators, allowing for relationships to be tested and implementation issues to be vetted and refined. Second, the district has made clear to its staff that it is moving forward with the implementation of its MMTES, giving the pilot year added weight and importance. However, the uniqueness of LAUSD and its IIP also restricts the generalizability of our findings. To begin with, the IIP consisted primarily of volunteers—a self-selected group of experienced, mostly elementary teachers who, according to both administrators from case study sites and AGT data, were particularly hard working and high performing (Strunk, Weinstein, & Makkonen, 2013). In addition, although we are able to study the relationships between measures of teacher effectiveness in a system as it is being implemented, and not in a research setting, the IIP

was not a high-stakes endeavor (i.e., attached to promotion or retention consequences), and the tools under study were still being revised during the study window. Given this context, we might expect that the relationships between AGT and observation-based measures during the IIP may not be a true reflection of what will be found in the full-scale roll out of the EGDC. However, since the majority of MMTES being implemented across the country are in their early pilot stages, LAUSD's experience provides an opportunity to assess relationships between measures of effectiveness in a context similar to that which many districts currently face. Nonetheless, it is with these constraints that we address the research questions outlined above.

The remainder of the paper proceeds as follows: The next section provides a brief review of the relatively recent body of work that examines the relationship between observation- and test score-based measures of teacher effectiveness. After that, we provide background on LAUSD's IIP of the EGDC and the unique context in which the EGDC is situated. Then we describe the data used in our analyses. We next present our findings from each of the research questions outlined above and discuss our results relative to those from similar studies in both experimental and practice-based settings. Last, we conclude with implications for districts and states that are implementing MMTES.

Summary of Related Literature

There has been a marked increase in studies that examine the relationships between value-added measures of teacher effectiveness and measures of effectiveness that are based on observations of classroom practice. We consider these studies in two groups: those that analyze the relationships between observation-based measures and VAMs in a purely research context (e.g., Grossman, Loeb, Cohen, & Wyckoff, 2013; Hill, Kapitula, & Umland, 2011; Kane & Staiger, 2012; and Mihaly, McCaffrey, Staiger, & Lockwood, 2013) and those that assess these relationships as measured in the implementation of systems in practice (e.g., Kane, Taylor, Tyler, & Wooten, 2011; Sartain et al., 2011; TDOE, 2012). These studies vary in their implementation context (research versus practice), as well as in the observational measures employed (those that examine just the classroom-based aspects of observational measures and those that explore the whole measure) and in the methods used to estimate relationships between observational and test-based measures of teacher effectiveness.

Together, the studies conducted in controlled research settings have found low to moderate correlations between VAMs and observation-based measures of teacher effectiveness. Using samples that varied in size from just 24 teachers (Grossman et al., 2013; Hill et al., 2011) to over 3,000 teachers in seven districts across the country (the Measures of Effective Teaching [MET] project) (Kane & Staiger, 2012; Mihaly et al., 2013), these studies tested multiple observational instruments and relied on expert reviewers to observe live and videotaped lessons. Simple and disattenuated correlations between the scores from the instructional standards of observation-based measures and "underlying" value-added measures of teacher effectiveness ranged in the MET study from 0.11 to 0.28 in reading or English Language Arts (ELA) and from 0.18 to 0.41 in math (Kane & Staiger, 2012; Mihaly et al., 2013). The authors of the MET studies worked to remove elements of bias from these relationships by assessing correlations between separate class sections from the same school year and between teachers' observational scores and their underlying value-added from the prior year.

Other research has suggested that the quality of instruction can vary between teachers with high and low value-added scores. For example, applying the Protocol for Language Arts Teaching Observation (PLATO), Grossman and colleagues (2013) identified systematic differences between

the observed instructional practices within 12 pairs of ELA teachers (who were matched based on their ranking in either high or low value-added quartiles), with the high value-added teachers earning significantly higher (p=.03) ratings on the PLATO domain related to Explicit Strategy Instruction. From this, the authors concluded that Explicit Strategy Instruction "appears to distinguish the more effective teachers in our sample," explaining that such teachers tend to provide students with very structured and specific ways to approach ELA activities and tasks, making "visible the often invisible processes requisite for successful, sophisticated literary analysis, reading comprehension, or writing" (p. 460). There is also some evidence that instruction can vary between math teachers with high and low value-added scores. Hill and colleagues (2011) rated the instruction of 24 purposively sampled middle school math teachers using the Mathematical Quality of Instruction (MQI) instrument (which monitors, for example, how explanations, representations, precise language, and mathematical generalizations are developed in the lesson). After controlling for the characteristics of each teacher's students, Hill and colleagues found a partial correlation of .52 ($p<.05$) between teachers' MQI ratings and their value-added scores, which the authors felt was indicative of "a teacher quality 'signal' in the scores" (p. 813).

It has been difficult to study the relationships between observational measures in practice (pilot or otherwise) and student test-based outcomes given that most MMTES are relatively new. Recent examinations of relationships in practice have emerged from Chicago and Tennessee, and several prior research studies have examined the Cincinnati Teacher Evaluation System (TES). Cincinnati has the longest-standing MMTES (launched in the 2000–2001 school year), and accordingly the most research has been done on this system, which is based (as many new MMTES are) on Charlotte Danielson's *Framework for Teaching* and accompanying observation protocol (Heneman & Milanowski, 2003; Holtzapple, 2003; Kane et al. 2011; Milanowski, 2004a, 2004b; Odden, 2004; Taylor & Tyler, 2012; Tyler, Taylor, Kane, & Wooten, 2010; Weisberg et al., 2009). Together, the work from Cincinnati finds that observation-based measures and value-added or student-achievement-based measures of teacher effectiveness are positively correlated, such that teachers who score well on one measure are more likely to be considered more effective according to the other.

The most recent work that examines these relationships in Cincinnati indicates that TES evaluators award higher observational ratings, at least on the two TES domains that can be directly observed in classroom settings (domains 2 and 3), to teachers whose students experienced higher gains in student achievement (Kane et al., 2011). According to Kane and colleagues (2011), improving a teacher's *overall classroom practices* TES score[1] by one point was associated with a one-seventh of a standard deviation increase in reading achievement among that teacher's students, compared to one-tenth in math. Notably, this study uses only the data elements from observations of classroom-based instruction (classroom environment and teaching), and does not assess the relationships between the observation score from the entire observation rubric (which consists of four domains) and value-added measures of teacher effectiveness or measures of student achievement growth. Although this provides information about the ability of the TES to reflect teacher effectiveness as measured by VAMs (and similarly, the ability of VAMs to reflect teacher effectiveness as measured by domains 2 and 3 of the TES), the ratings used in the study are not those that the teachers and principals receive as indicative of teachers' effectiveness based on the overall observation rubric.

[1] This score was calculated as the teacher's average score across the eight standards under TES Domain 2: Creating an Environment for Learning (3 standards) and TES Domain 3: Teaching for Learning (5 standards); prior to calculating this overall classroom practices score, the authors first averaged scores across evaluator observations for each individual practice and then averaged the individual practice scores within each standard (Kane et al., 2011).

Two recent initiatives have been studied in their early phases, and offer particular insight into this work. In 2008 the Chicago Public Schools (CPS) launched the Excellence in Teaching Pilot, an effort that was very similar to LAUSD's IIP. In their two-year study of this pilot effort, Sartain and colleagues (2011) also focused primarily on the classroom environment and instruction domains of Chicago's observation framework, for reasons similar to Kane and colleagues' (2011) in Cincinnati. For both math and reading teachers, Sartain and colleagues found statistically-significant relationships between teachers' observation ratings for the ten components in these two domains and their average value-added scores; teachers with the lowest observation ratings tended to have the lowest value-added scores and value-added scores tended to increase as ratings increased.[2] This "consistent correlation" suggested to the authors that the observation ratings "are a valid measure of teacher practice" (p. 56).

Tennessee implemented a statewide educator evaluation system for the 2011–2012 school year as part of its commitment under the federal Race to the Top grant competition. Under Tennessee's program, districts can use different observation instruments but must include in their evaluation systems value-added measures and observation-based measures of effectiveness.[3] According to the state's evaluation of the first year of implementation (TDOE, 2012, p. 33), each of the state's approved observation models "experienced alignment issues when taking into account student performance"—that is, the distributions of teachers' value-added and observation scores (each of which used scales ranging from 1 to 5) differed considerably, with observation ratings tending to cluster more in the upper end of the scoring range. For example, one county's state-approved observation instrument classified 96% of teachers in one of the top two performance levels, despite the fact that 19% of the teachers in the county scored in the lowest value-added level.

Overall these studies have found that observational and student test-based measures of teacher performance tend to be positively (but moderately) correlated, with a common explanation for these low to moderate correlations being that each measure contains distinct information about the teacher's underlying effectiveness. Certain empirical issues have arisen across these studies, however, in particular related to concerns about leniency among raters (Sartain et al., 2011; Sawchuck, 2013; Taylor & Tyler, 2012), potential nonlinearities between the two types of measures (Kane et al., 2011; TDOE, 2012), and the potential for biased assignment of students to teachers. Other literature has separately questioned the reliability of value-added and observational assessments of teachers. Several studies have found substantial year-to-year variability in teachers' VAM estimates (e.g., Aaronson, Barrow & Sander, 2007; Corcoran, 2010; McCaffrey et al., 2009; Newton et al., 2010), while other studies have found that observation scores vary considerably by rater and from lesson to lesson (e.g., Ho & Kane, 2013; Kane & Staiger, 2012; Rowan et al., 2013). These empirical concerns, although clearly important, are generally beyond the scope of this study. We do not seek to develop or refine any observational or student test-based measure of teacher effectiveness. Rather, evidence from this study is intended to help researchers and policymakers

---

[2] Sartain and colleagues (2011) used a simple regression model to assess this relationship. For each of the ten component ratings, the teacher's average value-added score was regressed on the teacher's rating (suppressing the intercept and using dummy variables for each of the four possible ratings). The authors then used an omnibus F-test to assess whether the ratings explained a significant portion of the variation in value-added, reporting (via their table A7) the F-statistics and p-values for the ten components, with separate tables for math and reading teachers. In reading, the test statistics indicated that seven of the ten component/value-added relationships were statistically significant at the .01 level, two were significant at the .05 level, and one was significant at the .10 level. In math, nine of the ten component/value-added relationships were statistically significant at the .01 level and one was significant at the .05 level.

[3] The majority of districts used the National Institute for Excellence in Teaching's evaluation model. The Tennessee State Board of Education also approved the use of three alternative observational measures in 12 municipal school districts across the state.

understand how these measures may operate in practice, when digested by practitioners implementing new MMTES, as opposed to in an experimental setting. Ultimately, despite potential psychometric issues with the measures involved, current federal, state and local policy requires their use. As a result, it is important to understand how these measures operate in practice.

This practice-focused lens has implications for the methods we use in our work. Specifically, because we are most interested in how teachers, principals and district administrators can and will use the information they receive from MMTES, we deviate from the studies discussed above in that we assess the relationship between VAMs and the overall practice-based measure of effectiveness (across all four domains) rather than just focusing on the classroom-based portions of the observation protocol. In addition, unlike in the MET study, we do not view within-year variation in participating teachers' observation scores as error. In fact, one of the core principles of most districts' MMTES reforms is that growth and development will occur during the evaluation year. Moreover, districts implementing MMTES will make personnel decisions and target support and development opportunities based on the actual scores given to teachers from observations at different times of the year, with lessons focused on different content and potentially conducted by different observers. In our conclusion section, we highlight difficulties districts may encounter when choosing the observation(s) to use in the final measure, how to address instances when teachers are only rated by one observer, and what to consider when teachers are rated on less than a complete set of observational elements.

Background on LAUSD's Initial Implementation Phase

The issues with most status quo systems of teacher evaluation also exist in LAUSD. In 2009, The New Teacher Project (TNTP) found that 99% of teachers in LAUSD received a "meets standard" rating, yet only 64% of teachers or principals reported that the evaluation provided them with sufficient information and strategies to help them improve teachers' practice (TNTP, 2009). To address deficiencies with the evaluation system and other issues pertaining to educator effectiveness, and based in part on recommendations from a district-stakeholder taskforce on the issue, the school board directed the district to develop a new system of teacher evaluation and support.

The resulting evaluation system, called the Educator Growth and Development Cycle (EGDC), was developed in the 2010–2011 school year, piloted in the 2011–2012 school year, and was originally intended to be taken to scale in the 2012–2013 school year. Upon scale-up, the EGDC was intended to include multiple measures of teacher effectiveness, including: (1) classroom observations of teacher practice by a site administrator and a second (external/off-site) observer using protocols aligned with the LAUSD Teaching and Learning Framework (adapted from Charlotte Danielson's *Framework for Teaching*); (2) stakeholder feedback surveys of students and parents; (3) teacher-, grade- and subject-level and school-wide value-added measures of teachers' contribution to student achievement on standardized test scores (AGT); and (4) a measure of teachers' contribution to the school community.

In 2010–2011, LAUSD worked with a group of internal and external partners to develop the LAUSD Teaching and Learning Framework (TLF). The TLF was intended to establish a common definition of effective teaching practice across the district and to serve as the cornerstone of teacher growth and development within the EGDC. Observation rubrics and templates for lesson designs, teacher self-assessments, and individual growth plans were also developed based on the TLF to be part of the EGDC. In addition, LAUSD worked with the Value-Added Research Center (VARC) to generate value-added measures of teacher performance (AGT). In the 2011–2012 school year, LAUSD began its Initial Implementation Phase (IIP) of the EGDC, piloting the use of teachers'

individual growth planning activities, the classroom observation cycle based on the new TLF-aligned protocols (with two observation cycles per year and two observers in each cycle), and the online platform for teachers and administrators to report observation notes and ratings. The IIP also piloted stakeholder surveys of students and parents for a subset of the participating teachers.[4] In addition, the district provided school-wide and individual AGT scores to all teachers in the district in tested grades and subjects.[5]

Data and Analysis Samples

The pilot sample consisted of 371 teachers in approximately 100 schools who were rated by any observer in any cycle. One hundred twenty-five site administrators (principals and assistant principals) served as primary raters, and 210 individuals were trained as second observers, 167 of whom entered a rating for a participating IIP teacher. This study primarily relies on two sets of data, both collected by LAUSD during the IIP year. The first dataset includes participating teachers' ratings by primary and secondary observers across TLF elements. The second includes teachers' AGT scores and assigned "level" in each subject/grade combination in which they had tested students. We do not include the stakeholder surveys because the sample sizes for these piloted measures were quite small, and the overlapping sample of teachers who piloted stakeholder surveys and had AGT or TLF scores was very small.

LAUSD Teaching and Learning Framework (TLF) Observation Ratings

LAUSD's TLF includes five standards (Planning and Preparation, Classroom Environment, Delivery of Instruction, Additional Professional Responsibilities, and Professional Growth),[6] 19 components (two to five per standard), and 63 elements (two to four per component). LAUSD's TLF-based rubric for rating teacher performance includes four rating levels (Ineffective, Developing, Effective, or Highly Effective), with descriptors defining performance on each element at each level. Based on feedback from teachers and administrators during the IIP, the district narrowed the original 63 elements into 19 "focus" elements on which observers were asked to rate participating IIP teachers during the 2011–2012 pilot year. These 19 focus elements are clustered within seven focus components. They are: 1d) Planning & Preparation: Designing coherent instruction; 1e) Planning & Preparation: Designing student assessment; 2b) Classroom Environment: Establishing a culture for learning; 3b) Delivery of Instruction: Using questioning and discussion techniques; 3c) Delivery of Instruction: Structures to engage students in learning; 3d) Delivery of Instruction: Using assessment in instruction to advance student learning; and, 5a) Professional Growth: Reflecting on practice.

Participating IIP teachers were to be rated in two separate observations (once during observation cycle 1, from October 2011 to February 2012, and then again in observation cycle 2, from March to May 2012) by his or her two observers. The observation data analyzed for this study include, for each of the 19 TLF focus elements, the teacher's rating (coded one to four), the observation cycle (coded one or two), and the rater who provided the score (primary or external). We calculate a number of measures of teacher effectiveness using teachers' TLF rankings: (1) *Overall*

[4] The district piloted its stakeholder (parent and student) surveys for a subset of teachers during the 2011–2012 school year. However, given low response rates and low initial sample sizes, we do not include these measures in our analyses.
[5] Tested grades and subjects refers to teachers who taught in a grade that was tested (second through eleventh grades), and in a subject that was tested. In LAUSD, VARC helped the district to generate VAMs for teachers who taught ELA, math, social studies and science classes.
[6] During the IIP, LAUSD decided to exclude Standard 4: Additional Professional Responsibilities. As a result, teachers were not rated on any components within this standard.

Average Score: The mean rating across all 19 focus elements, separately for each cycle and overall; (2) *Average Score by TLF Standard:* The mean rating for each TLF standard (four assessed) as the average of all available focus element scores within that standard, separately for each cycle and overall; and, (3) *Average Score by TLF Component:* The mean rating for each TLF component with focus elements (seven assessed), as the average of all available focus element scores within that component, separately for each cycle and overall. It is important to note that observers did not record the lesson type they observed, whether it was subject specific (i.e., math, ELA, science, etc.), or the kind of instruction they observed. This will be relevant as we attempt to assess the relationships between observation-based measures of effectiveness and value-added measures in math and/or ELA.

Table 1
Number of Teachers in IIP Rated on Focus Elements by Cycle and Observer

	Cycle 1				Cycle 2				Cycle 1 & Cycle 2			
Observer	P	S	B	E	P	S	B	E	P	S	B	E
Overall (4 standards, 7 components, 19 FEs)												
Rated on at least 1 FE	325	181	158	48	141	121	98	294	240	93	72	271
Rated on all 19 FEs	164	76	50	190	126	61	33	154	88	36	16	117
Rated on at least 1 FE per component	193	102	67	228	172	75	46	201	118	47	24	152
Rated on at least one FE per standard	212	109	77	244	183	81	53	211	128	52	28	163
Standard 1(2 components, 5 FEs)												
Rated on at least 1 FE	289	146	121	314	248	110	85	273	205	77	54	238
Rated on at least 1 FE per component	243	130	101	272	207	89	62	234	154	60	34	189
Rated on all 5 FEs	235	125	94	266	194	83	57	220	143	53	29	176
Standard 2 (1 component, 2 FEs)												
Rated on at least 1 FE	296	170	145	321	242	112	89	265	206	84	65	234
Rated on both FEs	275	151	126	300	211	100	74	237	178	72	51	207
Standard 3 (3 components, 10 FEs)												
Rated on at least 1 FE	321	179	153	347	260	119	94	285	230	91	70	263
Rated on at least 1 FE per component	287	166	137	314	237	112	86	255	200	81	60	223
Rated on all 10 FEs	216	113	81	248	159	84	57	186	124	51	27	156
Standard 5 (1 component, 2 FEs)												
Rated on at least 1 FE	221	114	81	254	187	85	55	217	135	52	28	171
Rated on both FEs	209	109	75	243	178	80	49	209	127	50	25	163

Note: P = Primary Observer; S = Secondary Observer; B= Both Primary and Secondary Observers; E = Either Primary or Secondary Observers.

Although in theory we should be able to calculate these measures by averaging across both observers in both cycles or in each cycle, this is not always feasible. This is because, as outlined in Table 1, teachers were rarely rated on all 19 focus elements by both observers in both cycles. During the first observation cycle (October 2011 to February 2012, shown in the first vertical panel of Table 1), 164 participating teachers (44% of pilot teachers) were rated on all 19 focus elements by their primary observer, and 76 (20%) were rated on all focus elements by their secondary observer, while only 50 (13%) of these teachers were comprehensively rated by both observers. The numbers were even lower during the second observation cycle (second vertical panel in Table 1), and only 16

teachers (4%) were rated on all focus elements by both observers in both cycles (third vertical panel in Table 1). Moreover, although the intent was for both observers to rate teachers on all focus elements in each cycle, this did not always occur.[7] To that end, only 72 teachers (19%) were rated *at all* (on at least 1 focus element) by both observers in both cycles, and only 158 and 98 (43 and 26%) were rated *at all* by both observers in Cycle 1 and Cycle 2, respectively.[8] Later in this section we will discuss how we address these inconsistencies by generating four analysis samples of IIP teachers based on the frequency with which they were rated on focus elements.

Academic Growth over Time (AGT) Scores

The Value-Added Research Center out of the University of Wisconsin generated value-added measures of teacher effectiveness for LAUSD for all teachers in grades and subjects covered by California Standards Tests (CSTs). VARC generated one-year and three-year average AGT scores for the teachers whenever feasible. For the LAUSD teacher-level model, AGT was measured in math in third through eighth grades, ELA in third through ninth grades, and the secondary level subjects Algebra I, Algebra II, Geometry, Biology, Chemistry, Physics, Integrated Science I, Science Grade 8, US History, and World History. CST scale scores were normalized (across the district's students) to have a mean of 0 and a standard deviation of 1, and only students continuously enrolled in the same school from the statewide school census date in October through the date of statewide testing were included. VARC's analysis included students with a posttest and pretest in consecutive grades in the same subject who could be assigned to a school, classroom, and teacher for that subject (VARC, 2011).

In general, the AGT model measures the "classroom effect"—the contribution a teacher made to his or her students' average CST achievement, controlling for prior student achievement in both math and ELA and a range of student- and classroom-level characteristics. At the elementary level, students' normalized CST scores are regressed on the student's prior achievement in the same subject and in another subject (math in the ELA model, ELA in the math model), vectors of student characteristics (gender, race, English learner and disability status, free- and reduced-price lunch status) and classroom characteristics (averages of the student characteristics), along with a vector of teacher indicators (which are, effectively, the teacher's/classroom's "value-add"). The AGT model for secondary level subjects yields estimates for teachers whose students took different pretests in prior years, and the secondary AGT model includes a term to control for average differences in posttest scores between students who took different pretests. For political and communication reasons, all AGT results provided by LAUSD are standardized to center around the number three such that the average teacher receives a "3" on his or her AGT.

Each teacher in the covered subjects receives an overall single- and three-year AGT score (as available). Results are produced for each grade taught—provided the grade has at least ten students, typically—and there is also an aggregate teacher measure that encompasses all of the grades taught by the teacher (VARC, 2011).[9] In addition, LAUSD provides teachers with their rating level, based

---

[7] Please see Strunk, Weinstein, and Makkonen (2013) for greater detail on implementation challenges faced by the district and the ongoing adjustments the district made to the intervention in response to these challenges and learning.

[8] It is possible that mid-course adjustments to the pilot that allowed the two observers to discuss their ratings and enter them together contributed to the decreased frequency with which participants had separate ratings entered by both observers over the course of the pilot.

[9] To construct these aggregate measures, VARC relies on individual grade-level scores in ELA and Math for grades 3-8, ultimately excluding teachers' estimates in Algebra 1, Algebra 2, and Geometry. Because this unnecessarily limits our sample to teachers in only grades 3-8, we replicate VARC's aggregate measure in both ELA and math, including teachers' estimates in Algebra 1, Algebra 2, and Geometry. In both math and ELA our one- and three-year aggregate measures are almost perfectly correlated with VARC's original measures (r=0.99) although our sample size is larger in

on their AGT scores and corresponding standard errors of measurement, on a five-level scale. These rankings are: Far Above Average, Above Average, Within Average, Below Average and Far Below Average. Teachers receive this information about their AGT scores for each subject/grade combination they teach. For our analyses, we use both the continuous AGT score given to teachers for each grade-subject combination and the five-level AGT measure based on all of the teacher's students in a subject (regardless of grade).[10]

Table 2
Summary Statistics for Overall AGT Measures; All Teachers in IIP and in LAUSD

	All Teachers w/2011-12 AGT					All Teachers in IIP					
	N	mean	sd	min	max	N	mean	sd	min	max	p-value
ELA 1 yr AGT	6747	3.009	0.874	-0.210	8.706	146	3.087	0.963	0.332	5.967	0.273
Math 1 yr AGT	7014	3.031	0.920	-0.513	8.986	141	3.280	0.977	0.574	6.137	0.001
ELA 3 yr AGT	7289	3.009	0.724	-0.210	7.173	157	3.148	0.779	0.960	5.340	0.015
Math 3 yr AGT	7445	3.020	0.809	0.472	8.225	148	3.265	0.899	0.607	5.510	0.000

Note. This table provides summary statistics of all continuous AGT measures for all teachers in LAUSD with 2011–2012 AGT and all teachers who participated in the IIP. Tests of group mean difference between IIP participants and all other teachers in LAUSD with 2011–2012 AGT were also performed and the results are provided in the p-value column.

Tables 2 and 3 show the descriptive statistics for these AGT measures for the district as a whole, as well as for all teachers who participated in the IIP. As expected, we see that the mean is around three (as noted above, AGTs were re-centered to three) and there is a substantial range in scores for the set of LAUSD teachers who have math and ELA AGT scores, ranging from a minimum just under zero to a high of approximately nine. However, this range is reduced in the set of teachers who participated in the IIP and who have AGT scores. Minimum scores in the IIP sample range from 0.62 to 1.62 of a standard deviation above the minimum scores in the full district sample and maximum scores are 2.53 to 4.10 standard deviations lower in the IIP sample than in the full district population. As is shown in Table 2, the mean in the IIP sample is just above three, suggesting that the pilot teachers are slightly (but not much) more effective on average, as measured by AGT, than other teachers in LAUSD. This is statistically significant for Math 1-year AGT ($p < 0.001$), ELA 3-year AGT ($p < 0.05$) and Math 3-year AGT ($p<0.001$).

As we will discuss in greater detail in the Results section, Table 3 indicates that very few teachers—in the district as a whole and in the IIP sample—score at the tails of the AGT distribution (Far Above Average or Far Below Average). This is the case regardless of the VAM type (1-year or 3-year, ELA or math). For example, only 1.5% of teachers are given a "Far Below Average" 1-year ELA rating, and only 2.3% of teachers are given a "Far Above Average" rating for 1-year ELA AGT. Because so few teachers score in extreme categories, we perform some of our analyses collapsing the five-level categories to three levels (Above and Far Above Average, Within Average, and Below and Far Below Average).

Table 3
Summary Statistics for AGT Rankings; All Teachers in IIP and in LAUSD

math. All analyses presented here rely on our aggregate measures. We also run the analyses using the aggregate measures provided by VARC, and our results remain consistent.

[10] VARC also generated AGT science and social science scores for a subset of teachers in the district. However, given the small samples of such teachers, we do not include these measures in our analyses.

	All Teachers w/2011-12 AGT			All Teachers in IIP			
	N	mean	sd	N	mean	sd	p-value
ELA 1 year % Far below avg	6747	0.015	0.120	146	0.027	0.164	0.189
ELA 1 year % Below avg	6747	0.117	0.321	146	0.082	0.276	0.187
ELA 1 year % Within avg	6747	0.725	0.447	146	0.719	0.451	0.872
ELA 1 year % Above avg	6747	0.121	0.326	146	0.130	0.338	0.722
ELA 1 year % Far above avg	6747	0.023	0.149	146	0.041	0.199	0.135
ELA 3 year % Far below avg	7289	0.009	0.093	157	0.025	0.158	0.023
ELA 3 year % Below avg	7289	0.145	0.352	157	0.102	0.303	0.121
ELA 3 year % Within avg	7289	0.640	0.480	157	0.592	0.493	0.207
ELA 3 year % Above avg	7289	0.180	0.384	157	0.236	0.426	0.066
ELA 3 year % Far above avg	7289	0.026	0.160	157	0.045	0.207	0.145
Math 1 year % Far below avg	7014	0.024	0.153	141	0.021	0.145	0.826
Math 1 year % Below avg	7014	0.200	0.400	141	0.156	0.364	0.186
Math 1 year % Within avg	7014	0.527	0.499	141	0.482	0.501	0.288
Math 1 year % Above avg	7014	0.198	0.399	141	0.255	0.438	0.087
Math 1 year % Far above avg	7014	0.051	0.220	141	0.085	0.280	0.060
Math 3 year % Far below avg	7445	0.020	0.139	148	0.027	0.163	0.520
Math 3 year % Below avg	7445	0.213	0.409	148	0.142	0.350	0.033
Math 3 year % Within avg	7445	0.472	0.499	148	0.405	0.493	0.103
Math 3 year % Above avg	7445	0.244	0.429	148	0.311	0.464	0.054
Math 3 year % Far above avg	7445	0.052	0.223	148	0.115	0.320	0.001

Note. This table provides summary statistics of all non-continuous AGT measures for all teachers in LAUSD with 2011–2012 AGT and all teachers who participated in the IIP. Tests of group mean difference between IIP participants and all other teachers in LAUSD with 2011–2012 AGT were also performed and the results are provided in the p-value column.

Analysis Samples

Given the constraints with the observation data, we clearly do not want to limit analyses to just those 16 teachers who were observed on all 19 focus elements in both observation cycles by both observers. Not only would this make it nearly impossible to detect significant relationships, but it is likely that many districts implementing rubric-based observations will face similar problems of "attrition" unless districts require observers to rate teachers on all focus elements and reduce the number of focus elements on which teachers must be rated. Instead, we generate four separate study samples: (1) the 210 participating teachers who have an ELA or math AGT score and were rated on at least one focus element rating by at least one observer in at least one cycle (Sample 1); (2) the 194 teachers who have an ELA or math AGT score and were rated on at least ten focus elements by at least one observer in at least one cycle (Sample 2); (3) the 166 teachers with math or ELA AGT scores who are rated on at least one focus element by at least one observer in both observation cycles; and (4) the 99 teachers who have an ELA or math AGT score and were rated on at least one focus element by a primary *and* secondary observer in either Cycle 1 or Cycle 2. In addition, we examine these four groups of participating teachers in the first and second cycles separately, as well as combined. Tables 4 and 5 provide the summary statistics for AGT measures for each sample.

Table 6 shows the sample sizes, means and standard deviations for teachers' observational score overall, and by standard and component for each sample of teachers in the combined observation cycles and in each cycle separately. As shown, means and standard deviations do not differ substantially for the overall, standard or component measures across samples. This indicates that observational ratings are, on average, quite similar regardless of the sample used. We note, however, that Sample 4 teachers—those rated by both a primary and secondary observer in the same

cycle—consistently score slightly lower on average, although the variance around the mean is consistent between groups. However, this difference is generally on the order of 0.02 or 0.03 points on the four-point scale, indicating that this difference is substantively small. Given these similarities, we focus our analyses moving forward on Sample 2 for ease of interpretation, and because it is the sample that most closely resembles how LAUSD (and many other districts) will likely move forward with the MMTES ratings.[11]

A slightly different story emerges, however, when we analyze observational scores across cycles. We find that there is frequently a relatively large difference between observation ratings in Cycle 1 versus Cycle 2. For instance, we see a quarter of a standard deviation increase in overall TLF rating, on average, between Cycles 1 and 2. Observation ratings by standard increase, on average, 0.10 to 0.25 standard deviations between the two cycles, and we see increases between cycles of similar magnitudes for all focus components within Standards 1, 2 and 3. This is not necessarily surprising—increases between Cycles 1 and 2 may indicate that teachers are improving in their practice over the course of the year. However, because professional development was not the focus of the IIP, LAUSD did not provide participants with professional development opportunities aligned to their observation ratings. Ratings increases between cycles, then, may suggest that the mere act of participating in the IIP improved teachers practice, or that teachers and observers became more familiar with the TLF and/or the EGDC process by the second observation cycle. Alternately, it may be that the differences between ratings in Cycles 1 and 2 indicate observer bias—observers may want to show increases in teacher effectiveness between cycles, or they may be more lenient in Cycle 2 than in Cycle 1. We cannot assess the degree to which any of these rationales holds true in this paper.[12] Nonetheless, the rating discrepancy evidence between cycles has implications for the choices districts make about how to weight or aggregate observations across cycles. Although previous studies have examined the teacher's average observational score across multiple cycles during the year, districts may want to acknowledge teachers' progress by weighing Cycle 2 more heavily, or by simply focusing their attention on Cycle 2 scores alone. Following the previous literature but also taking into account these findings, we focus our analyses on the relationship between AGT and the observational measure averaged across both cycles and in Cycle 2 alone.

Tables 4 and 5 show that there are relatively small differences in mean AGT across the samples of teachers. Average AGT scores are highest in Sample 2 for the ELA and overall AGT scores, but math scores are slightly higher, on average, in Sample 3. Given these differences, we run all analyses on Sample 3, as well. In addition, average AGT scores are lowest in Sample 4, and greater proportions of teachers in Sample 4 have AGT scores in the Far Below Average and Below Average categories, echoing the earlier finding that teachers in Sample 4 are slightly less effective as measured both by TLF rankings and by AGT scores. Given these differences, we run all analyses on Sample 4, as well. We find that all results presented below are substantively the same when we use Samples 3 and 4 as opposed to Sample 2, and are available upon request from the authors. Although substantively the same, analyses using Sample 4 differ from those in Sample 2 more so than do the

[11] Specifically, moving forward LAUSD has chosen to reduce the number of observers to one (the primary observer). In addition, LAUSD is refining the TLF, but likely will not reduce the number of focus elements drastically. Given this, we believe that Sample 2 most closely resembles the future TLF rating scheme in the EGDC.

[12] That teachers improve their practice between cycles is contradicted by evidence (available upon request), that participating teachers' AGT does not significantly increase between 2010–2011 (the year before the IIP) and 2011–2012 (the year of the IIP), controlling for a set of teacher and school covariates. These latter findings are in line with those reported in Taylor and Tyler (2012), who show that teachers VAMs do no significantly improve over the course of their participation in the Cincinnati TES.

results from Sample 3. However, Sample 4 results rely on very small sample sizes, which may help to partially explain these minor differences (Ns range from approximately 50 to 80 teachers).

Table 4

Summary Statistics for Overall AGT Measures; Sample 1, Sample 2, Sample 3, Sample 4

	Sample 1					Sample 2					Sample 3					Sample 4				
	N	Mean	sd	min	max	N	mean	sd	min	max	N	mean	sd	min	Max	N	mean	sd	min	Max
ELA 1yr AGT	140	3.102	0.977	0.332	5.967	130	3.127	0.985	0.332	5.967	108	3.093	0.981	0.552	5.967	74	3.046	1.003	0.552	5.472
Math 1yr AGT	133	3.304	0.985	0.574	6.137	123	3.327	1.006	0.574	6.137	98	3.359	1.051	0.574	6.137	68	3.296	1.084	0.574	5.757
ELA 3yr AGT	149	3.162	0.793	0.960	5.340	137	3.189	0.762	1.411	5.340	112	3.128	0.795	0.960	5.340	79	3.133	0.816	0.993	4.978
Math 3yr AGT	139	3.299	0.905	0.607	5.510	128	3.317	0.895	0.607	5.510	101	3.371	0.956	0.607	5.510	72	3.296	0.960	0.607	5.510

Table 5

Summary Statistics for AGT Rankings; Sample 1, Sample 2, Sample 3, Sample 4

	Sample 1			Sample 2			Sample 3			Sample 4		
	N	Mean	sd	N	mean	sd	N	Mean	sd	N	mean	sd
ELA 1 year % Far below avg	140	0.029	0.17	130	0.023	0.151	108	0.028	0.17	74	0.041	0.2
ELA 1 year % Below avg	140	0.079	0.27	130	0.077	0.268	108	0.093	0.29	74	0.108	0.31
ELA 1 year % Within avg	140	0.714	0.45	130	0.708	0.457	108	0.694	0.46	74	0.649	0.48
ELA 1 year % Above avg	140	0.136	0.34	130	0.146	0.355	108	0.148	0.36	74	0.176	0.38
ELA 1 year % Far above avg	140	0.043	0.2	130	0.046	0.211	108	0.037	0.19	74	0.027	0.16
ELA 3 year % Far below avg	149	0.027	0.16	137	0.015	0.12	112	0.036	0.19	79	0.038	0.19
ELA 3 year % Below avg	149	0.101	0.3	137	0.095	0.294	112	0.107	0.31	79	0.101	0.3
ELA 3 year % Within avg	149	0.577	0.5	137	0.606	0.49	112	0.571	0.5	79	0.544	0.5
ELA 3 year % Above avg	149	0.248	0.43	137	0.234	0.425	112	0.25	0.44	79	0.278	0.45
ELA 3 year % Far above avg	149	0.047	0.21	137	0.051	0.221	112	0.036	0.19	79	0.038	0.19
Math 1 year % Far below avg	133	0.015	0.12	123	0.016	0.127	98	0.02	0.14	68	0.029	0.17
Math 1 year % Below avg	133	0.158	0.37	123	0.154	0.363	98	0.163	0.37	68	0.176	0.38
Math 1 year % Within avg	133	0.474	0.5	123	0.463	0.501	98	0.429	0.5	68	0.397	0.49
Math 1 year % Above avg	133	0.263	0.44	123	0.268	0.445	98	0.276	0.45	68	0.279	0.45
Math 1 year % Far above avg	133	0.09	0.29	123	0.098	0.298	98	0.112	0.32	68	0.118	0.32
Math 3 year % Far below avg	139	0.022	0.15	128	0.016	0.125	101	0.03	0.17	72	0.014	0.12
Math 3 year % Below avg	139	0.129	0.34	128	0.141	0.349	101	0.119	0.33	72	0.167	0.38
Math 3 year % Within avg	139	0.403	0.49	128	0.398	0.492	101	0.386	0.49	72	0.375	0.49
Math 3 year % Above avg	139	0.324	0.47	128	0.328	0.471	101	0.317	0.47	72	0.319	0.47
Math 3 year % Far above avg	139	0.122	0.33	128	0.117	0.323	101	0.149	0.36	72	0.125	0.33

Table 6

Average Overall, Standard and Component Observation Ratings by Analysis Sample and Observation Cycle

		Sample 1			Sample 2			Sample 3			Sample 4		
	Cycle	N	mean	sd	N	mean	sd	N	mean	sd	N	mean	sd
Overall	both	210	2.73	0.4	191	2.74	0.39	150	2.72	0.38	99	2.70	0.42
	1	194	2.69	0.42	178	2.69	0.41	150	2.67	0.41	97	2.65	0.43
	2	166	2.79	0.42	153	2.80	0.41	150	2.79	0.43	83	2.79	0.44
Standard 1	both	199	2.82	0.42	190	2.82	0.42	142	2.81	0.38	96	2.80	0.43
	1	177	2.77	0.44	172	2.77	0.44	135	2.75	0.43	93	2.73	0.46
	2	153	2.89	0.43	148	2.88	0.43	138	2.88	0.42	79	2.87	0.43
Standard 2	both	198	2.76	0.46	191	2.77	0.44	143	2.76	0.41	98	2.73	0.47
	1	177	2.75	0.48	172	2.76	0.46	136	2.74	0.45	93	2.71	0.47
	2	149	2.8	0.49	146	2.80	0.49	135	2.81	0.49	78	2.77	0.52
Standard 3	both	210	2.66	0.42	191	2.66	0.42	150	2.64	0.42	99	2.62	0.44
	1	194	2.62	0.46	178	2.62	0.45	150	2.59	0.45	97	2.57	0.46
	2	163	2.69	0.46	151	2.70	0.46	147	2.70	0.47	80	2.69	0.46
Standard 5	both	175	2.91	0.54	175	2.91	0.54	133	2.88	0.54	93	2.85	0.57
	1	147	2.86	0.58	147	2.86	0.58	115	2.83	0.58	78	2.82	0.61
	2	124	2.98	0.63	124	2.98	0.63	114	2.98	0.64	70	2.96	0.64
Component 1d	both	199	2.86	0.41	190	2.86	0.42	142	2.84	0.38	96	2.84	0.43
	1	177	2.8	0.45	172	2.80	0.45	135	2.78	0.44	93	2.78	0.47
	2	153	2.92	0.42	148	2.92	0.41	138	2.91	0.41	79	2.90	0.42
Component 1e	both	183	2.67	0.58	181	2.66	0.58	135	2.67	0.53	95	2.62	0.55
	1	157	2.63	0.6	155	2.62	0.6	122	2.61	0.58	88	2.57	0.63
	2	135	2.72	0.65	135	2.72	0.65	122	2.73	0.63	72	2.69	0.60
Component 2b	both	198	2.76	0.46	191	2.77	0.44	143	2.76	0.41	98	2.73	0.47
	1	177	2.75	0.48	172	2.76	0.46	136	2.74	0.45	93	2.71	0.47
	2	149	2.8	0.49	146	2.80	0.49	135	2.81	0.49	78	2.77	0.52
Component 3b	both	210	2.61	0.5	191	2.62	0.5	150	2.59	0.49	99	2.57	0.51
	1	193	2.58	0.55	178	2.58	0.55	149	2.54	0.55	97	2.54	0.54
	2	159	2.65	0.54	149	2.66	0.54	143	2.67	0.55	79	2.61	0.54
Component 3c	both	201	2.76	0.45	189	2.76	0.45	147	2.75	0.45	98	2.73	0.48
	1	180	2.7	0.48	174	2.71	0.47	141	2.67	0.47	93	2.64	0.50
	2	154	2.84	0.49	144	2.84	0.5	139	2.85	0.50	79	2.85	0.50
Component 3d	both	198	2.64	0.46	188	2.64	0.44	146	2.60	0.43	98	2.58	0.44
	1	173	2.6	0.47	169	2.60	0.47	134	2.57	0.46	92	2.55	0.45
	2	154	2.66	0.52	146	2.66	0.52	141	2.65	0.51	79	2.62	0.49
Component 5a	both	175	2.91	0.54	175	2.91	0.54	133	2.88	0.54	93	2.85	0.57
	1	147	2.86	0.58	147	2.86	0.58	115	2.83	0.58	78	2.82	0.61
	2	124	2.2	1.13	124	2.20	1.13	114	2.21	1.13	70	2.96	0.64

Results and Discussion

Do Value-Added and Observational Measures of Teacher Effectiveness Provide Teachers and Principals with a Consistent Signal of Effectiveness?

To answer our first research question we regress teachers' AGT scores on their average observation rating. Rather than use teachers' average AGT scores when they teach more than one grade-subject combination, we use their individual grade-subject AGT scores and cluster our standard errors to the teacher level.

Table 7
AGT Scores Regressed on Observational Scores, Overall and by Standard

	ELA 1 year	ELA 3 year	Math 1 year	Math 3 year
Overall TLF	0.216**	0.142**	0.178*	0.149*
	(0.08)	(0.05)	(0.09)	(0.07)
R2-adjusted	0.045	0.033	0.027	0.023
N Observations	134	179	134	178
N Teachers	130	137	123	128
Standard 1	0.196*	0.129**	0.200*	0.157*
	(0.08)	(0.05)	(0.09)	(0.07)
R2-adjusted	0.027	0.023	0.035	0.025
N Observations	134	179	133	177
N Teachers	130	137	122	127
Standard 2	0.253**	0.176**	0.158+	0.098
	(0.08)	(0.06)	(0.09)	(0.09)
R2- adjusted	0.06	0.045	0.02	0.007
N Observations	134	179	134	178
N Teachers	130	137	123	128
Standard 3	0.195**	0.134**	0.156+	0.149*
	(0.07)	(0.05)	(0.08)	(0.07)
R2-adjusted	0.038	0.03	0.019	0.022
N Observations	134	179	134	178
N Teachers	130	137	123	128
Standard 5	0.201*	0.099+	0.188+	0.200*
	(0.09)	(0.06)	(0.11)	(0.09)
R2-adjusted	0.037	0.011	0.025	0.036
N Observations	124	166	124	163
N Teachers	120	127	113	117

Note: Regressions based on Sample 2 averaged across both cycles, + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

The top panel of Table 7 shows results from these regressions. Coefficients are shown in standard deviation units, such that they can be interpreted as the proportion of a standard deviation difference in AGT score associated with a standard deviation increase in TLF score. For all regressions we use teachers in Sample 2, with TLF scores averaged across both cycles as our main specification, but provide an identical table showing the same regressions on teachers looking at just Cycle 2 observational scores in the appendix (Appendix Table A1).[13] We discuss differences in relationships when using the average across cycles as opposed to Cycle 2 scores in our final section.

Table 7 shows that there are significant positive relationships between the overall observational score and all AGT measures, indicating that, on average, teachers and principals who receive a VAM- and an observation-based measure of effectiveness under LAUSD's MMTES will obtain similar signals about their effectiveness. However, the low to moderate size of the association means that, while the two measures give the same general signal about effectiveness on average, they may also provide teachers and administrators with unique information about their levels of effectiveness. Alternately, it may mean that there is some "noise" in one or both measures of effectiveness that tells teachers and administrators little or nothing about teacher effectiveness. If this is the case, even if both the observational and test-score-based constructs reflect the same underlying aspects of teacher effectiveness, the noise in the measures would ensure that the two could never be perfectly correlated.

The strongest relationship is between 1-year ELA AGT and overall observational score, with a one standard deviation increase in overall observational score associated with a 0.22 standard deviation increase in 1-year ELA AGT. To roughly compare this to findings from the Cincinnati TES, in which Kane and his colleagues (2011) found that a one-unit increase in TES ranking based on domains 2 and 3 is associated with a one-seventh standard deviation increase in reading achievement and a one-tenth increase in math achievement, we find that a one unit increase in overall TLF rank (SD=0.39) is associated with approximately a little over one-half of a standard deviation increase in 1-year ELA (SD = 0.985) and a little under one-half of a standard deviation increase in 1-year math AGT (SD=1.006).

Although these relationships are substantively meaningful, the observation measure explains little of the variation in AGT scores, with a maximum of 4.5%.[14] Interestingly, while the relationships between math AGT and overall observational scores are of about the same magnitude (a one standard deviation increase in the overall observation measure is associated with a 0.18 standard deviation increase in 1-year Math AGT), they are significant at the $p<.05$ level, and the observational score explains less of the variation in the math AGT measures. Our finding regarding the small (but significant) proportion of variance in value-added that is explained by component-level observation rating scores is in line with what Sartain and colleagues (2011) found in Chicago's pilot program.[15] Such similarities are not surprising, as both contexts involve pilot implementations of new Danielson Framework-based observation measures in large urban school districts.

---

[13] Results from identical analyses using Sample 3 provide substantively the same results, and are available upon request from the authors. Some relationships lose significance, but this appears to be largely due to reductions in the analysis sample size.

[14] This is likely an underestimate of the true R-squared value. Because teachers' VAM scores contain some amount of estimation error, the R-squared will necessarily understate the raters' skills in identifying practices related to value-add. Nonetheless, even if we assume a high proportion of the AGT score is noise, the maximum R-squared will still be relatively low. See Aaronson, Barrow and Sander (2007) and Koedel and Betts (2007) for more detail.

[15] Although the analytic samples and pilot implementations differed substantially, both studies regressed pilot teachers' math and ELA value-added scores for the year on their Danielson Framework-based scores for components 2b, 3b, 3c and 3d and tended to identify statistically-significant ($p<.10$) relationships (except in the case of the relationship between

For easier comparison with results presented in the MET study (Kane & Staiger, 2012), Table 8 shows these relationships as correlations between each AGT measure and the overall observation rating in Sample 2, in the average of Cycles 1 and 2 as well as Cycle 2 alone. Again using the specification from both cycles, we see correlations that range from 0.18 (3-year Math) to 0.24 (1-year ELA). Kane and Staiger (2012) found correlations between the Danielson Framework-based observational rating on domains 2 and 3 and math VAMs of 0.19 and of 0.11 with ELA VAMs (correlated between separate classes in the same year, with nearly identical relationships between VAMs and observation ratings using prior-year VAMs with current-year observation scores). However, the correlations we find between math AGT and observational measures from the IIP are substantially smaller than those reported in the final 2013 MET report (Mihaly et al., 2013). Although the correlations between ELA AGT and observation measures in LAUSD are of the same magnitude as those reported in Mihaly et al. (2013), we presume that the observed discrepancies in Math AGT and observational score correlations are due at least in part to Mihaly and colleagues' methodological choice to examine correlations between observation ratings and the "stable component" of VAMs, as opposed to our analysis of correlations between AGT as reported and observation ratings from the IIP, as well as our use of the full observation rating as opposed to just scores from domains 2 and 3.

Table 8
Correlations between AGT Scores and Overall Observational Scores

Correlations between TLF Score and:	Both cycles	Cycle 2
ELA - 1 year	0.24	0.29
Math - 1 year	0.19	0.18
ELA - 3 year	0.22	0.28
Math - 3 year	0.18	0.19

Table 9
Average Overall Observation Rating by 5-Level and 3-Level AGT

	5-level AGT					3-level AGT		
	1	2	3	4	5	1/2	3	4/5
ELA - 1 year	2.42	2.71	2.75	2.84	3.04+	2.64	2.75	2.89+
ELA - 3 year	3.07	2.50	2.78	2.84	3.01	2.58+	2.78	2.87
Math - 1 year	2.85	2.72	2.70	2.81	2.98+	2.74	2.70	2.86+
Math - 3 year	3.15	2.57	2.73	2.83	2.89	2.66	2.73	2.85+

Note: All values based on Sample 2 averaged across both cycles. T-tests for 5-level AGT include FBA vs. all other AGT levels and FAA vs. all other AGT levels. T-tests for 3-Level AGT include BA/FBA vs. all other AGT levels and AA/FAA vs. all other AGT levels.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Because simply being given a number—a value-add "score"—is likely difficult for teachers and administrators (and most non-statisticians) to interpret, LAUSD teachers receive a report that contains their AGT score (bounded by a confidence interval), as well as their placement within the five AGT levels. We hypothesize that teachers and administrators pay more attention to their AGT level, which is easily interpretable, than they do to their raw AGT score. If this is indeed the case, it

one-year math value-added and component 3c in the LAUSD pilot, which was not statistically significant at the .10 level).

is also important for districts implementing MMTES to consider the relationships between teacher practices as measured by observation-based rubrics and the VAM-associated level of effectiveness.

The left panel of Table 9 shows the average observational score for teachers at each of the five AGT levels. We test for significant differences between the observational scores of teachers who are ranked in the lowest category (Far Below Average, labeled 1 in Table 9) and all other teachers, as well as between teachers who are ranked in the highest category (Far Above Average, labeled 5 in Table 9) and all other teachers. We find that observation ratings do not consistently map to AGT scores across AGT levels. Specifically, teachers who score at the Far Below Average level (1) on AGT do not consistently have lower overall observational scores than teachers at the Below Average (2), or even the Within Average (3) or Above Average (4) levels. This is especially the case when examining the 3-year AGT measures. For instance, teachers in the 3-year ELA Far Below Average group have a mean overall observational score of 3.07, which is higher than the average overall observation-based scores for any other group. We see a similar pattern for the 3-year Math AGT score. This is likely due in part to the very small proportions of teachers who fall into the extreme categories. (As we showed earlier in Table 5, only approximately 1.5% of teachers in Sample 2 score in the 3-year ELA and 3-year Math Far Below Average categories.) Nonetheless, the non-linear relationships shown in Table 9 indicate that teachers who score in the lower tail on the AGT level measure may receive inconsistent messages about their effectiveness from their VAM level and TLF score.

Next we collapse the five-level AGT measure into three levels, shown in the right panel of Table 9. Once we do this, we see evidence of a more linear relationship between AGT and overall observational score across all AGT levels. We also see that teachers who score in the top AGT category in the three-level measure perform significantly better ($p<.10$) on the observational ratings than teachers in the Within Average and Below/Far Below Average categories across all AGT measures, except the 3-year ELA AGT. However, we find relatively little evidence that teachers in the combined Below/Far Below Average category perform significantly differently than others on the overall observation rating. One exception here is in the relationship between overall observation rating and 3-year ELA AGT, where we find those in the Far Below Average/Below Average category have significantly lower observational scores ($p<.10$) than other teachers.

In short, we find that teachers who receive both an AGT score and an observational TLF score will receive somewhat consistent signals from both measures of effectiveness. However, the two measures do not map perfectly to each other, and the associations between the two measures are only low to moderate, indicating that teachers may well receive different indications of their effectiveness from each measure.

Do Specific Classroom Practices Measured by a District-Generated Observation Rubric Capture Differences in Teacher Effectiveness, as Measured by Value-Added Scores?

We next turn to our second research question, which asks if specific instructional practices captured by classroom observations are associated with value-added measures of teacher effectiveness in LAUSD. This is particularly important for teachers and administrators, who will be working to understand observation-based feedback about their specific practices in light of their AGT data. We first examine this by again regressing each AGT measure on the four TLF focus standards. These results can be found in the bottom four panels of Table 7. We see clear evidence that, again, teachers' ratings on the TLF-based observations tend to be significantly associated with their AGT, although these associations are still only in the low to moderate range. Upon closer inspection a number of interesting findings emerge.

First, we turn our attention to the relationships between AGT and teacher performance on Standards 2 and 3, the standards most frequently explored in the previous literature (e.g., Kane et al., 2011 and Sartain et al., 2011). Teachers' Standard 2 ratings are less clearly associated with math AGT scores (the positive relationships are only significant at the $p<.10$ level for 1-year Math AGT and are not significant for 3-year Math AGT). Teachers were only rated on one focus component within Standard 2 (Establishing a Culture for Learning), and within that there were only two focus elements (Expectations for Learning, and Achievement and Student Ownership of Their Work). It is possible that observers had a more difficult time assessing these skills as they relate to math instruction, which would result in noisy estimates of the relationship between math AGT and Standard 2 ratings. This may be due to insufficient training of raters to assess these skills in the context of math lessons, or to the difficulty of applying the training to math instruction. Alternately, it is possible that the skills captured by Standard 2 are less reflected in teachers' abilities to raise student achievement in math than in ELA. Given that we do not know (and neither did the district) what type of lesson was being observed (math, ELA or another subject entirely), it is difficult to determine which, if any, of these possible explanations is correct.

We also find positive and significant relationships between AGT scores and ratings on Standard 3, although again these relationships are not as significant in math as in ELA. Standard 3 captures teachers' Delivery of Instruction, and in the IIP consisted of three focus components with a total of ten focus elements. The first two focus components (3b: Using Questioning and Discussion Techniques and 3c: Structures to Engage Students in Learning), map most closely to the "Explicit Strategy" and "Student Engagement" domains from the different observational frameworks that Grossman and colleagues (2013) found to be most significantly associated with value-added measures of teacher effectiveness. Table 10 provides the results from our analyses that regress AGT scores on the specific focus components. The row that corresponds with Component 3b shows that both math and ELA AGT remain significantly associated with teachers' ratings on "Using Questioning and Discussion Techniques," whereas the row below it (which corresponds with Component 3c) indicates that math AGT is not significantly related to teachers' ratings on "Structures to Engage Student Learning." In addition, the third focus component included in Standard 3 (3d: Using Assessment in Instruction to Advance Student Learning) is also less significantly related to math than to ELA AGT scores. Although we cannot test this empirically, it is again possible that observers have a more difficult time rating these skills and practices for math teachers or for math lessons than they do for ELA teachers and lessons, or that teachers in LAUSD who are very effective in raising their students math test scores are not necessarily those who excel in the skills measured by TLF Standard 3, and vice versa.

We also see evidence that teachers' AGT scores are positively and significantly associated with their ratings in Standard 1 (Planning and Preparation) and Standard 5 (Professional Growth). This is particularly important in the context of our study, as these results indicate that teachers receive similar signals from their VAM scores as they do from their scores on domains that other studies with a validity perspective have not considered. Standard 1 overall and Components 1d and 1e all seem relatively well associated with AGT scores in both math and ELA. A teacher's rating on Component 1e (Designing Student Assessment, assessed via focus element 1e4: Analysis and Use of Assessment Data for Planning) in particular is significantly associated with their math AGT scores. This may indicate that using and analyzing assessment data to plan classes and activities translates particularly strongly to teachers' ability to increase students' math achievement.

Table 10
AGT Scores Regressed on Observation Component Scores

	ELA 1 year	ELA 3 year	Math 1 year	Math 3 year
Component 1d	0.191*	0.129**	0.178*	0.131+
	(0.09)	(0.05)	(0.09)	(0.07)
R2-adjusted	0.026	0.022	0.027	0.017
N Observations	134	179	133	177
N Teachers	130	137	122	127
Component 1e	0.216**	0.151**	0.254**	0.236**
	(0.08)	(0.06)	(0.09)	(0.08)
R2-adjusted	0.04	0.032	0.052	0.053
N Observations	127	170	125	167
N Teachers	123	130	114	119
Component 2b	0.253**	0.176**	0.158+	0.098
	(0.08)	(0.06)	(0.09)	(0.09)
R2- adjusted	0.06	0.045	0.02	0.007
N Observations	134	179	134	178
N Teachers	130	137	123	128
Component 3b	0.178*	0.132*	0.157+	0.169*
	(0.08)	(0.06)	(0.09)	(0.07)
R2-adjusted	0.029	0.029	0.02	0.032
N Observations	134	179	134	178
N Teachers	130	137	123	128
Component 3c	0.177*	0.122**	0.138	0.121
	(0.07)	(0.05)	(0.09)	(0.08)
R2-adjusted	0.028	0.022	0.013	0.013
N Observations	133	178	132	175
N Teachers	129	136	121	126
Component 3d	0.189*	0.128*	0.144+	0.111
	(0.08)	(0.05)	(0.08)	(0.08)
R2-adjusted	0.035	0.027	0.013	0.008
N Observations	133	178	132	175
N Teachers	129	136	121	126
Component 5a	0.201*	0.099+	0.188+	0.200*
	(0.09)	(0.06)	(0.11)	(0.09)
R2-adjusted	0.037	0.011	0.025	0.036
N Observations	124	166	124	163
N Teachers	120	127	113	117

Note: Regressions based on Sample 2 averaged across both cycles, + $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

In the same way that it is important for districts implementing MMTES to understand if teachers' AGT levels (and not just their scores) discriminate between the overall instructional quality measured by observation rubrics, we believe that it is informative to examine the degree to which specific elements of teachers' instructional practice vary across AGT levels. Table 11 shows teachers' average observation ratings for each TLF standard and component in the collapsed 3-level AGT. Our first observation is that, generally, teachers who are the most adept at increasing student achievement (Above or Far Above Average) are rated significantly better than other teachers in the specific practices captured by the observation measure. To the contrary, we do not see a great deal of evidence that teachers who are Far Below or Below Average in AGT scores are rated significantly lower on TLF standards or components. This may be due to a number of factors.

Table 11
Observation Standard and Component Ratings by 3-Level AGT

Standard 1	1/2	3	4/5	Standard 3	1/2	3	4/5
ELA 1yr	2.67+	2.84	2.90	ELA 1yr	2.61	2.65	2.84*
ELA 3yr	2.71	2.85	2.94	ELA 3yr	2.49+	2.70	2.81+
Math 1yr	2.81	2.80	2.93	Math 1yr	2.68	2.60	2.78*
Math 3yr	2.70	2.84	2.93+	Math 3yr	2.58	2.62	2.77+
Component 1d				Component 3b			
ELA 1yr	2.69+	2.90	2.91	ELA 1yr	2.63	2.63	2.81
ELA 3yr	2.76	2.90	2.96	ELA 3yr	2.40*	2.71	2.77
Math 1yr	2.85	2.83	2.96	Math 1yr	2.67	2.55	2.78*
Math 3yr	2.75	2.87	2.96+	Math 3yr	2.54	2.57	2.76+
Component 1e				Component 3c			
ELA 1yr	2.50	2.65	2.86+	ELA 1yr	2.71	2.77	2.93+
ELA 3yr	2.46	2.66	2.87*	ELA 3yr	2.63	2.79	2.93+
Math 1yr	2.56	2.65	2.82+	Math 1yr	2.81	2.68	2.90*
Math 3yr	2.43*	2.70	2.79+	Math 3yr	2.69	2.69	2.90*
Standard 2 (Component 2b)				Component 3d			
ELA 1yr	2.70	2.80	2.95+	ELA 1yr	2.51	2.59	2.79*
ELA 3yr	2.60*	2.87	2.89	ELA 3yr	2.46	2.63	2.75
Math 1yr	2.76	2.75	2.92+	Math 1yr	2.60	2.59	2.73+
Math 3yr	2.71	2.77	2.90*	Math 3yr	2.54	2.64	2.71
				Standard 5 (Component 5a)			
				ELA 1yr	2.65+	2.92	3.05
				ELA 3yr	2.71	2.94	3.02
				Math 1yr	2.85	2.89	3.09*
				Math 3yr	2.88	2.90	3.03

Note: All values based on Sample 2 averaged across both cycles. T-tests for 3-Level AGT include BA/FBA vs. all other AGT levels and AA/FAA vs. all other AGT levels. + $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

First, it is possible that raters are hesitant to give low scores to teachers on components, which we see given the relatively low proportions of teachers who receive low observational scores overall and by standard/component. As a result, even teachers who are truly inadequate in their practice may be given scores on their observations that do not reflect their true instructional

capacities. This may reflect a form of "rater leniency," similar to that found in earlier work by Sartain and colleagues (2011) in Chicago's pilot and noted by Taylor and Tyler (2012) in Cincinnati. Second, it might be that the trainings provided to raters on how to observe and score teachers did not provide observers with the capacity to differentiate between inadequate and average levels of practice. Third, it is possible that the observation rubric is not capable of capturing true instructional differences at the low end of the effectiveness spectrum. Of course, we also do not dismiss the possibility that teachers who are particularly high-ranking on the TLF elements and standards use different instructional practices than the average teacher, but that the remainder of teachers' practice is simply not very associated with student achievement outcomes among the teachers assessed in LAUSD's IIP. Again, given data limitations, we are unable to investigate the extent to which any of these rationales holds true.

When we examine the specific practices assessed by the various standards and components, we find evidence that there is a strong relationship between the instructional process measured by TLF Standard 3 and student test performance. This seems particularly driven by Component 3c (Structures to Engage Student Learning), again paralleling earlier findings from Grossman et al. (2013), that show that the CLASS "Student Engagement" domain captures elements of teaching that are associated with increases in student achievement. TLF Component 1e (Designing Student Assessment) also appears to discriminate relatively well between teachers who most greatly contribute to student achievement on standardized tests and other teachers.

Together, the evidence presented in Tables 4, 7 and 8 indicate that specific instructional practices captured by the observation measure as used in the initial implementation of the EGDC do modestly reflect teachers' abilities to improve student achievement, particularly in ELA and especially in the current year (as opposed to a three-year average AGT measure). More specifically, these results suggest that teachers who are particularly effective in raising student achievement on standardized tests are rated significantly higher on TLF Standard 3: Delivery of Instruction.

Limitations and Future Work

The intent of this paper is to examine the implications of using a new standards-based multiple measure teacher evaluation system in practice rather than in an experimental setting. While we point out a number of important implications from this work, it is important to highlight some of the limitations to our analyses and offer suggestions for future work on district use of MMTES. The first set of limitations arise from the pilot nature of the intervention. As we acknowledge earlier in the paper, although all tested teachers in the district received AGT scores during our study, the pilot sample of teachers who received TLF observation scores consists of a set of volunteers. This has implications for the generalizability of our results. In addition, the sample of pilot teachers with both AGT and TLF scores is relatively small, lessening the precision with which we can assess relationships between the two measures. The pilot itself also brings some limitations in that LAUSD was pointedly learning from the pilot implementation, and as a result elements of the intervention were not completely finalized and subject to change. Last, the pilot intervention, like many MMTES implemented across the country, did not require observers to record the lesson type they were observing. This prevents us from knowing if the relationships under study arise from observations of lessons in the same subject as the AGT measures, or from different subjects. If we assume that teacher practice and effectiveness differs across subjects, then this may result in an underestimate of the relationships between test-based and observation-based measures of effectiveness. The second set of limitations arises from the necessity for this analysis of testing multiple relationships. There is always a chance for Type I error in assessing statistical significance, and this is of course heightened

when there are so many regressions and correlations using the same data. As a result, we consider our findings to be suggestive and not causal.

Practical Implications for the Implementation of MMTES

It is reassuring in many ways that the findings we present above are so similar to those found in other studies of districts' use of MMTES and in studies that examine these measures in research contexts. In particular, these results indicate that it is possible for the results from early implementations of observation-based measures of effectiveness to have similar relationships with value-added measures as found in research settings and/or in long-standing evaluation systems. However, it is the *differences* in our results from the previous studies that may be particularly useful to district administrators and policymakers implementing MMTES. In the discussion that follows, we highlight aspects of our findings that should be considered by districts and policymakers as they implement MMTES.

Overall, Unadjusted Observation-Based Measures and VAMs Provide Teachers with a Modestly Consistent, but not Identical, Signal of Effectiveness.

The results discussed above use teachers' average TLF score across all four domains assessed during the IIP. This is considerably different from previous work that has relied on just the scores from the two domains that can be best observed during actual classroom instruction (Kane et al., 2011; Kane & Staiger, 2012). In addition, whereas some other work has used multiple techniques to explore relationships between observational measures and teachers' "true" effectiveness (Ho & Kane, 2013; Mihaly et al., 2013) and made adjustments to avoid concerns about omitted variable and other forms of bias in VAM scores (Chetty, Friedman, & Rockoff, 2011; Rockoff & Speroni, 2011), we take the viewpoint of the teacher and administrator receiving multiple pieces of data from the new MMTES. We find that teachers' VAM-based measures of effectiveness are, on average, somewhat reflective of their observation-based measure of effectiveness. This means that teachers rated on both elements from their MMTES will receive a modestly consistent signal about their effectiveness: if they are scored as effective on one, they will likely be scored as effective on the other.

However, the two measures are only correlated at relatively weak or moderate levels, indicating that the two measures still provide different information to teachers and their administrators. Theoretically, the VAM-based measure tells teachers their effectiveness in improving student achievement scores in a given year (1-year VAMs) and across three years (3-year VAMs), whereas the observation-based score provides teachers with information about their specific teaching practices. The overlap between the two measures indicates that these practices are modestly associated with, but not entirely indicative of their abilities to improve test scores. As such, teachers and principals should still be able to learn something about teachers' practice, strengths, and areas for growth from the observation-based measure that they do not learn from the VAMs, and vice-versa.

However, it should be noted that TLF scores explain less than 5% of the variation in AGT scores. This means that the far majority of the test-based score is *not* explained by the TLF. In other words, as districts implement MMTES, teachers will receive some consistent signal from the two measures in an evaluation system, but likely will also receive differing assessments of their effectiveness. This may lead to confusion and frustration. This study cannot tell us what does explain the remainder of the AGT score, and future research that assesses the validity and use of MMTES should further explore these relationships.

A District-Generated Observational Protocol Like the TLF Can Provide Information about Teachers' Instructional Practice that is Consistent to Varying Degrees with VAMs.

Ideally, the more granular information that stems from observational protocol like the TLF should help teachers and principals determine in which areas of practice teachers are strong and where they need to improve. We find that the TLF measure provides information about teachers' practice that is both modestly consistent with the information they receive from VAMs, but also varies by standard/component. This may indicate that some parts of the TLF better capture teachers' abilities to raise student test scores, whereas others may be more reflective of teacher practices that are not as helpful in raising test scores. This information may help the district determine which instructional practices are most directly related to student achievement so that they can develop and implement professional development that seeks to improve teacher practice in ways that positively impact student performance.

Our results also demonstrate that evidence from the relationship between specific components of instructional practice outlined by observation-based protocols and VAMs may be beneficial for districts as they seek to determine the specific focus elements to include in the classroom observation measure of their MMTES. As evidenced by LAUSD, time and manpower concerns forced the district to identify a subset of 19 focus elements from the original 63 elements included in the TLF rubric during the IIP, and the district has further reduced the number of focus elements in later iterations of the reform. Using information on the relationships between these elements and VAM results may help districts like LAUSD better determine which focus elements to include in the MMTES.

There May be a Role for Both Three-Year and One-Year VAM Scores in MMTES.

The differences we find between relationships between TLF scores and one-year versus three-year VAMs (with higher correlations with one-year VAMs) suggest that there may also be a role in MMTES for both sets of VAMs. Although VAMs are more stable and are generally thought to more reliably measure teacher effectiveness when they are generated from at least three years of data, we find that the three-year AGT scores are less associated with observation measures than are the one-year AGT scores in LAUSD. This may be because the one-year AGT scores only measure teachers' performance in the *same* year as the observation measure, whereas the three-year AGT reflects teachers' performance over a number of years. This suggests that, although three-year VAMs likely capture a teacher's "true" effectiveness better than one-year VAMs, single-year VAMs may better capture *this* year's performance. Although some variability in value-added measures of teacher performance across years is simply "noise" (e.g., resulting from differences in testing conditions and other factors outside of teachers' control), the stronger associations between results from one-year VAMs and same-year observational measures may suggest that the measurement of such year-to-year variability is substantively important. Districts may choose to use three-year VAMs in MMTES for summative purposes because they are more stable over time and less likely to shift significantly based on a specific class or situation in a given year. However, the one-year VAM score may provide teachers with some relevant information about their practices in the most recent year, and in conjunction with observation data may be useful for formative reasons.

Observation-Based Measures of Effectiveness May Be More Useful in Differentiating between Top and Average Performers than Distinguishing Low Performers.

Although, on average, the two measures explored in this study give modestly consistent signals of teacher effectiveness, it appears that teachers who score in the top two levels of a five-level value-added ranking are more likely to have higher observation scores (overall and by specific

practices), whereas teachers who score in the bottom two levels are no more likely to have lower observation scores. This may be in part due to the fact that observers in the IIP year are relatively unlikely to give teachers a low score on their observations, overall or within any particular standard. Given possible "rater leniency" (similar to that shown in earlier work by Sartain and colleagues, 2011, and Taylor and Tyler, 2012), it appears that the TLF-based measure, as implemented during the IIP year, is better able to discriminate between teachers with above average AGT scores and other teachers than to distinguish teachers who fall below average in their AGT scores.

This may be a function of the TLF-based measure itself or it may reflect insufficient training of raters in how to discriminate between teacher (in)effectiveness at the lower end of the effectiveness distribution. Alternately, this may be a function of the sample of teachers and principals who participated in the IIP. Teachers volunteered or were asked to participate in the IIP, and principals reported that participating teachers were a stronger, more committed and hard-working group, on average, than other teachers in their schools. This is supported by the fact that the IIP samples' mean AGT scores were significantly higher than the overall sample of teachers in LAUSD. If this is the case, it may simply be that few teachers who participated in the IIP were low performers. However, we note that the AGT scores for pilot participants do not fully support this hypothesis, as there were actually slightly greater proportions of teachers who received Far Below Average AGT ratings in the IIP sample than in the district overall.

In addition, we show that the district's five-level AGT ratings are not as well-aligned with observation-based measures of performance as are the three-level AGT ratings created for this study, at least during the IIP year. This may be because so few teachers score at either tail of the AGT distribution, so outliers can easily sway the average observational scores for those groups of teachers. Nonetheless, it may be more helpful for administrators and teachers to consider the practices of teachers who score anywhere below or above the average range, as opposed to attempting to determine what practices specifically identify the teachers at the tails. This would, in effect, reduce the five-level VAM placement to three levels. However, this will not entirely solve the problem that emerges from the reality that VAMs and observational scores are only moderately aligned.

The Subject of the Lesson Observed May Matter for Signal Consistency

Results presented above show that TLF ratings are more strongly associated with ELA VAM results than with math VAM results. Given that the MET studies found stronger relationships between math VAM results and Danielson Framework observation ratings than were evident for ELA, it is somewhat surprising that our results show the opposite. We posit that this may have occurred because the majority of participants in the IIP were elementary school teachers, whereas the MET study included many more middle school teachers. Middle school teachers are more likely to specialize in their content areas—teaching either math *or* ELA, whereas elementary school teachers teach both topics. Similarly, administrators at the secondary level may be more adept at observing content-specific instruction as compared to primary school principals. To the extent that raters have a more difficult time discriminating between levels of teaching quality in math than in ELA among elementary school teachers, we may expect to see lower correlations between math AGT and observational scores in the IIP.[16] In addition, we do not know what subject lesson was being observed during the IIP (the data were not collected by the district). If the lessons observed focused mostly on literacy or a related topic, then theoretically math VAMs should not be as highly

[16] Research underlying the MQI has suggested that mathematical work that occurs in classrooms—particularly the presence of mathematical explanations and practices—is distinct from classroom climate, pedagogical style, or the use of more general instructional strategies (see, for example, Hill, Umland, Litke, & Kapitula, 2012).

correlated with the observation results as ELA VAMs.[17] Regardless, these findings suggest that districts may want to note the subject of the lesson type being observed and consider observing different kinds of lessons to ensure that teachers receive feedback on instruction in more than one subject (where applicable).

The Quantity of Observed Focus Elements and the Number of Observers May Be Less Important than Simply Observing Teachers

Our finding that means, standard deviations, and patterns in the analyses are substantially the same across all four samples of teachers has implications for the implementation of MMTES. Specifically, our first analysis sample consisted of the 210 teachers who had AGT scores and were rated on at least *one* focus element by at least *one* observer in *one* observation cycle. Sample 2 teachers had to have been rated on at least *ten* focus elements by at least *one* observer, again in *one* observation cycle. Sample 3 included teachers who were again rated on at least *one* focus element, but this time in *both* observation cycles. Sample 4, by contrast, included teachers who had been again rated on at least *one* focus element, but this time by *both* observers in at least *one* of the cycles. That the descriptive statistics for each of these samples, and the patterns and relationships observed in the data, remain substantively the same suggests that it may not matter whether teachers are rated on at least one, or at least ten, elements. Moreover, whether they are rated by one or both observers seems to matter little. One of the great difficulties with implementing MMTES is the feasibility of the systems. It takes administrators a great deal of time to score and tag teacher practice on multiple elements, and it is difficult to coordinate with multiple observers (see Strunk, Weinstein, & Makkonen, 2013, for more detail). The results provided in this paper suggest that districts might streamline the process.

The Observation(s) Included in the Final Observation Score Matters

Our results also have implications for the choice of observation (s) to use in summative measures of teacher effectiveness. We find that ratings differ between the first and second cycle, with teachers scoring higher across all TLF standards in Cycle 2 compared to Cycle 1. Districts should consider if they want to evaluate teachers on their "best effort" observation later in the year, when they have had the chance to improve, or if they want to average observations over both cycles to reflect a teacher's average level of effectiveness in a given year. However, if the higher observational scores in Cycle 2 reflect raters' bias, such that they expect to see teachers improve over the course of the year or they are hesitant to give low scores to teachers who have been working hard to improve over the course of the year, then capturing Cycle 2 scores may less reflect teachers' "best efforts" than raters' leniency.

Moreover, the decision of which observation(s) to incorporate into the final evaluation may shift the relationship between the observational score and the value-added measure of effectiveness. As is shown in Appendix Tables A1-A3, relationships between ratings on TLF Standard 5 and both ELA and math AGT increase in magnitude and significance when examining correlations in Cycle 2 as opposed to the average across cycles. (Standard 5 rates teachers' "Professional Growth.") This increase in the relationship between the two measures in Cycle 2 may indicate that observations of

[17] It is important to note that Charlotte Danielson's *Framework for Teaching* is designed to be a content-neutral observation instrument focusing only on observing instructional practices that one might expect to observe across content areas. Despite its content-neutral nature, observers using the Danielson Framework to rate instructional practice are themselves not necessarily content-neutral observers. Given this, we believe it is important for districts to collect data on the content of the lesson being observed in order to empirically test the extent to which the Danielson Framework as implemented by observers in practice is consistently applied across content areas.

professional growth may be noisy in the first cycle, before growth can be observed, and may be more accurate during the second cycle. This may indicate that districts should consider if specific standards and practices are better rated at different times of the year.

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. http://dx.doi.org/10.1086/508733

Anderson, J. (2012, February 20). States try to fix quirks in teacher evaluations. *New York Times*, p. A1.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87. http://dx.doi.org/10.1080/10627197.2012.715014

Brandt, C., Mathers, C., Oliva, M., Brown-Simms, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest Region* (Issues and Answers Report, REL 2007 –No. 030). Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, REL Midwest.

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

Corcoran, S. (2010). Can teachers be evaluated by their students' test scores? Should they be? (Report for the Education Policy for Action Series). Providence, RI: Annenberg Institute for School Reform at Brown University.

Donaldson, M. L. (2009). So long, Lake Wobegon? Using teacher evaluation to raise teacher quality. Center for American Progress.

Fleisher, L. (2012, September 10). Teacher grading off to uneven start. *Wall Street Journal*, p. A21.

Goldhaber, D., Brewer, D. & Anderson, D. (1999). A three-way error components analysis of educational productivity. *Education Economics, 7*, 199-208. http://dx.doi.org/10.1080/09645299900000018

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. *American Journal of Education, 119*(3), 445-470. http://dx.doi.org/10.1086/669901

Heneman, H. G. III, & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(2), 171-195.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794–831. http://dx.doi.org/10.3102/0002831210387916

Hill, H. C., Umland, K. U., Litke, E., & Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education, 118*(4), 489-519. http://dx.doi.org/10.1086/666380

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel (research paper). Measures of Effective Teaching (MET) project, Bill and Melinda Gates Foundation.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207-219. http://dx.doi.org/10.1007/s11092-005-2980-z

Jerald, C. (2012). Movin' it and improvin' it! Using both education strategies to increase teaching effectiveness. Washington, DC: Center for American Progress.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment (research paper). Measures of Effective Teaching (MET) project, Bill and Melinda Gates Foundation.

Kane, T. J. & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains (research paper). Measures of Effective Teaching (MET) project, Bill and Melinda Gates Foundation.

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. http://dx.doi.org/10.1353/jhr.2011.0010

Kauchak, D., Peterson, K., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19*, 32-37.

Koedel, C., & Betts, J. (2007). Re-Examining the Role of Teacher Quality in the Educational Production Function. Working Paper, University of Missouri.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. http://dx.doi.org/10.1162/edfp.2009.4.4.572

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching (research paper). Measures of Effective Teaching (MET) project, Bill and Melinda Gates Foundation.

Milanowski, A. (2004a). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53. http://dx.doi.org/10.1207/s15327930pje7904_3

Milanowski, A. (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 TC-04–02.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23).

Odden, A. (2004). Lessons Learned About Standards-Based Teacher Evaluation Systems. *Peabody Journal of Education, 79*(4), 126-137. http://dx.doi.org/10.1207/s15327930pje7904_7

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. http://dx.doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252. http://dx.doi.org/10.1257/0002828041302244

Rockoff, J. E., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18*(5), 687–696. http://dx.doi.org/10.1016/j.labeco.2011.02.004

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay and student achievement. *Quarterly Journal of Economics, 25*(1), 175–214. http://dx.doi.org/10.1162/qjec.2010.125.1.175

Rowan, B., Schilling, S. G., Spain, A., Bhandari, P., Berger, D., & Graves, J. (2013). Promoting high quality teacher evaluations in Michigan: Lessons from a pilot of educator effectiveness tools. Institute for Social Research, University of Michigan.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Consortium on Chicago School Research.

Sawchuk, S. (2013). Teachers' ratings still high despite new measures: Changes to evaluation systems yield only subtle differences. *Education Week, 32*(20), 1,18-19.

Strunk, K. O., Weinstein, T., & Makkonen, R. (2013). Understanding the Implementation of a Standards-Based Multiple Measure Teacher Evaluation Reform. Paper presented at March 2013 Association for Education Finance and Policy annual conference.

Taylor, E. S. & Tyler, J. H. (2012). The effect of evaluation on teacher performance. American *Economic Review 102*(7), 3628-3651. http://dx.doi.org/10.1257/aer.102.7.3628

Tennessee Department of Education. (TDOE 2012). Teacher evaluation in Tennessee: A report on year 1 implementation. Nashville, TN: Author.

The New Teacher Project (TNTP). (2009). Teacher hiring, transfer, and evaluation in Los Angeles Unified School District. Final Report.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using Student Performance Data to Identify Effective Classroom Practices. *American Economic Review: Papers & Proceedings, 100*(2), 256-260. http://dx.doi.org/10.1257/aer.100.2.256

Value-Added Research Center (VARC). (2011). Academic Growth over Time: Technical report on the LAUSD teacher-level model academic year 2010-2011. Los Angeles, CA: Los Angeles Unified School District.

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

Appendix

Table A1
AGT Scores Regressed on Observational Scores, Overall and by Standard

	ELA 1 year	ELA 3 year	Math 1 year	Math 3 year
Overall TLF	0.243**	0.162**	0.167+	0.168+
	(0.08)	(0.06)	(0.10)	(0.09)
R2-adjusted	0.071	0.047	0.019	0.023
N Observations	112	144	110	144
N Teachers	109	112	101	104
Standard 1	0.192*	0.116*	0.136	0.105
	(0.09)	(0.06)	(0.10)	(0.10)
R2-adjusted	0.035	0.017	0.008	0.004
N Observations	108	140	106	140
N Teachers	105	108	97	100
Standard 2	0.262**	0.162*	0.096	0.063
	(0.09)	(0.07)	(0.10)	(0.11)
R2-adjusted	0.075	0.041	0	-0.003
N Observations	107	137	106	138
N Teachers	104	107	97	100
Standard 3	0.214**	0.154**	0.127	0.159+
	(0.08)	(0.05)	(0.10)	(0.09)
R2-adjusted	0.055	0.044	0.007	0.021
N Observations	111	143	109	143
N Teachers	108	111	100	103
Standard 5	0.325**	0.176*	0.260*	0.241*
	(0.11)	(0.08)	(0.13)	(0.11)
R2-adjusted	0.108	0.044	0.053	0.055
N Observations	90	115	90	119
N Teachers	88	91	81	84

Note: Regressions based on Sample 2, Cycle 2, +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table A2
Average Overall Scores by 5-Level and 3-Level AGT – Sample 2 Cycle 2

	5-level AGT					3-level AGT		
	1	2	3	4	5	1/2	3	4/5
ELA 1 year	2.09*	2.65	2.83	2.85	3.12	2.55*	2.83	2.90
ELA 3 year	2.96	2.50	2.80	2.93	3.03	2.57*	2.80	2.94+
Math 1 year	2.93	2.78	2.71	2.92	2.98	2.83	2.71	2.94*
Math 3 year	3.17	2.61	2.78	2.88	2.90	2.70	2.78	2.89

Note: All values based on Sample 2 in cycle 2. T-tests for 5-level AGT include group 1 vs. all other AGT levels and group 5 vs. all other AGT levels. T-tests for 3-Level AGT include groups 1/2 vs. all other AGT levels and 4/5vs. all other AGT levels.
+ $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table A3
AGT Scores Regressed on Observation Component Scores, Sample 2 Cycle 2

	ELA 1 year	ELA 3 year	Math 1 year	Math 3 year
Component 1d	0.173+	0.112+	0.116	0.096
	(0.09)	(0.06)	(0.10)	(0.10)
R2-adjusted	0.025	0.014	0.003	0.002
N Observations	108	140	106	140
N Teachers	105	108	97	100
Component 1e	0.198*	0.111+	0.175+	0.103
	(0.08)	(0.06)	(0.10)	(0.11)
R2-adjusted	0.039	0.015	0.017	0.002
N Observations	99	127	96	127
N Teachers	96	99	87	90
Component 2b	0.262**	0.162*	0.096	0.063
	(0.09)	(0.07)	(0.10)	(0.11)
R2-adjusted	0.075	0.041	0	-0.003
N Observations	107	137	106	138
N Teachers	104	107	97	100
Component 3b	0.186*	0.144*	0.156	0.185+
	(0.08)	(0.07)	(0.12)	(0.10)
R2-adjusted	0.037	0.037	0.016	0.033
N Observations	109	139	107	139
N Teachers	106	109	98	101
Component 3c	0.187*	0.158**	0.085	0.109
	(0.08)	(0.06)	(0.10)	(0.10)
R2-adjusted	0.037	0.047	-0.003	0.006
N Observations	104	132	103	133
N Teachers	101	104	94	97
Component 3d	0.211**	0.133*	0.089	0.107
	(0.08)	(0.06)	(0.09)	(0.09)
R2-adjusted	0.05	0.028	-0.002	0.005
N Observations	108	140	107	141
N Teachers	105	108	98	101
Component 5a	0.325**	0.176*	0.260*	0.241*
	(0.11)	(0.08)	(0.13)	(0.11)
R2-adjusted	0.108	0.044	0.053	0.055
N Observations	90	115	90	119
N Teachers	88	91	81	84

Note: Regressions based on Sample 2, Cycle 2, +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table A4
Standard and Component Ratings by 3-Level AGT - Sample 2 Cycle 2

Standard 1	1/2	3	4/5	Standard 3	1/2	3	4/5
ELA 1yr	2.63*	2.93	2.90	ELA 1yr	2.47+	2.73	2.83
ELA 3yr	2.71	2.87	2.97	ELA 3yr	2.45*	2.70	2.88*
Math 1yr	2.95	2.83	3.00	Math 1yr	2.74	2.59	2.83*
Math 3yr	2.79	2.91	2.96	Math 3yr	2.62	2.66	2.79
Component 1d				Component 3b			
ELA 1yr	2.70*	2.99	2.89	ELA 1yr	2.45	2.73	2.77
ELA 3yr	2.79	2.94	2.97	ELA 3yr	2.33*	2.73	2.83
Math 1yr	3.00	2.87	3.03	Math 1yr	2.70	2.51	2.83*
Math 3yr	2.82	2.95	3.00	Math 3yr	2.54	2.59	2.77+
Component 1e				Component 3c			
ELA 1yr	2.45	2.71	2.90	ELA 1yr	2.70	2.88	2.97
ELA 3yr	2.59	2.60	2.97*	ELA 3yr	2.65	2.85	3.01+
Math 1yr	2.75	2.63	2.88	Math 1yr	2.92	2.71	2.99*
Math 3yr	2.64	2.74	2.79	Math 3yr	2.80	2.73	2.97*
Standard 2 (Component 2b)				Component 3d			
ELA 1yr	2.54*	2.88	2.94	ELA 1yr	2.31*	2.67	2.78
ELA 3yr	2.48**	2.89	2.94	ELA 3yr	2.39+	2.64	2.82*
Math 1yr	2.75	2.76	2.97*	Math 1yr	2.67	2.61	2.74
Math 3yr	2.70	2.76	2.95*	Math 3yr	2.54	2.69	2.71
				Standard 5 (Component 5a)			
				ELA 1yr	2.67+	3.03	3.17
				ELA 3yr	2.71	2.98	3.18+
				Math 1yr	2.97	2.87	3.27**
				Math 3yr	2.86	2.98	3.11

Note: All values based on Sample 2 in cycle 2. T-tests for 3-Level AGT include groups 1/2vs. all other AGT levels and groups 4/5vs. all other AGT levels. + $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

About the Author

Katharine O. Strunk
Associate Professor of Education and Policy
University of Southern California.
kstrunk@usc.edu
Dr. Katharine Strunk is an Associate Professor of Education and Policy at the University of California. Dr. Strunk's research focuses on issues related to education governance and teacher labor markets. She received her PhD in Administration and Policy Analysis and her MA in Economics from Stanford University.

Tracey L. Weinstein
Director of Policy and Innovation
StudentsFirst
tweinstein@studentsfirst.org
Dr. Tracey Weinstein is the Director of Policy and Innovation at StudentsFirst, a national advocacy organization that works on K-12 policy issues. Her research focuses on the intersection of district-level reform and the teacher labor market, specifically the implementation and effects of teacher quality initiatives on the composition of the teacher labor market.

Reino Makkonen
Senior Policy Associate
WestEd
rmakkon@wested.org
Dr. Reino Makkonen is a Senior Policy Associate at WestEd, where his research focuses on teacher workforce policy and school and district leadership, including examining current reforms in the assessment of teachers and school leaders. Makkonen received his PhD from the University of California Berkeley Graduate School of Education, with a focus on policy analysis and measurement.

education policy analysis archives

Volume 22 Number 100 November 10th, 2014 ISSN 1068-2341

Education Research Complete, ERIC, Education Full Text (H.W. Wilson), QUALIS A2 (Brazil), SCImago Journal Rank; SCOPUS, SOCOLAR (China).

Please contribute commentaries at http://epaa.info/wordpress/ and send errata notes to Gustavo E. Fischman fischman@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.

archivos analíticos de políticas educativas
consejo editorial

arquivos analíticos de políticas educativas
conselho editorial